Department of Physics

How Structural Modifications Affect the Regulation and Activation Mechanisms of CD44 and JAK2

Joni Vuorio

Doctoral thesis, to be presented for public examination with the permission of the Faculty of Science of the University of Helsinki, in Auditorium B123, Exactum building, on the 7th of June, 2021 at 16 o'clock.

University of Helsinki
Finland

Supervisor
Prof. Ilpo Vattulainen, University of Helsinki, Helsinki, Finland

Pre-examiners
Prof. Bert de Groot, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany
Prof. Lars Schäfer, Ruhr University, Bochum, Germany

Opponent
Prof. Roland Faller, University of California, Davis, USA

Custos
Prof. Ilpo Vattulainen, University of Helsinki, Helsinki, Finland

Contact information

Department of Physics, University of Helsinki
P.O. Box 68 (Pietari Kalmin katu 5)
FI-00014, Helsinki
Finland

ISBN 978-951-51-7300-3 (paperback)
ISBN 978-951-51-7301-0 (PDF)
Helsinki 2021
Unigrafia Oy

Abstract

Cell surface receptor proteins are important gatekeepers. They enable cells to react to their surroundings by receiving and propagating chemical signals through the cell membrane. Countless such reactions occur in the cells of our bodies every moment to keep us alive. Hence, understanding how these molecules operate helps us to improve our health and the quality of our life.

In this Thesis, we discuss the activation of two cell signaling-related proteins, CD44 and JAK2. CD44 is a cell surface receptor for a carbohydrate called hyaluronan. Through hyaluronan binding, CD44 is involved in various signaling cascades that regulate, *e.g.*, cell–cell interactions as well as cell differentiation, proliferation, and survival. JAK2 is a non-receptor tyrosine kinase — an intracellular signaling protein that binds to cytokine receptors, forming a receptor–JAK signaling complex. Through this interaction, JAK2 mediates central physiological functions, including haematopoiesis and immune response.

Despite their physiological significance, the understanding of these proteins is incomplete because their atomic-level operating principles have not yet been fully elucidated. Therefore, we use biomolecular simulations to shed light into the function and regulation of these proteins and their cognate molecules. In the case of CD44, we expand the current knowledge on the details of its hyaluronan binding. We also show how N-glycosylation of the receptor can modulate the ligand binding by altering its binding site preference. In the case of JAK2, we find how the activation of the signaling complex is controlled by intracellular dimerization and proper orientation through specific membrane binding.

Our work provides novel atomic-level insight into the functions and interactions of the studied proteins. The results can be useful in drug development — especially in the search for new drug binding sites, for example, at glycosylation, dimerization, or membrane binding interfaces of proteins. Finally, this work highlights the added value gained by bridging computer simulations with experimental techniques.

Tiivistelmä

Solupinnan reseptoriproteiinit ovat tärkeitä portinvartijoita. Ne vastaanottavat solunulkoisia signaaleja ja välittävät niitä eteenpäin käynnistämällä solujen sisäisiä kemiallisia kaskadeja. Lukemattomia tällaisia reaktioita tapahtuu kehomme soluissa joka hetki, jotta pysyisimme hengissä. Siksi näiden molekyylien ja niiden toiminnan ymmärtäminen on keskeistä esimerkiksi sairauksien hoidossa.

Tässä väitökirjassa käsittelemme kahden solun signalointiin liittyvän proteiinin, CD44:n ja JAK2:n, aktivaatiota. CD44 on luonnollisen hiilihydraattipolymeerin, hyaluronihapon, solupintareseptori. Hyaluronihapon sitoutumisen kautta CD44 osallistuu erilaisiin signalointikaskadeihin, jotka säätelevät esimerkiksi solujen välistä vuorovaikutusta sekä solujen kasvua, lisääntymistä ja eloonjäämistä. JAK2 puolestaan on tyrosiinikinaasi, joka ei kuitenkaan itse toimi reseptorina. Se on signalointiproteiini, joka sitoutuu sytokiinireseptoreihin muodostaen signalointikompleksin, johon kuuluu sekä JAK2 että reseptori. Tämän vuorovaikutuksen kautta JAK2 säätelee monia kehomme prosesseja, kuten hematopoieesia ja immuunivastetta.

Käsitys näiden proteiinien toimintaperiaatteista on edelleen puutteellista, sillä niiden molekyylitason toimintaa ei vielä täysin ymmärretä. Tämän vuoksi käytämme tietokonesimulaatioita valaisemaan kyseisten proteiinien toimintaa ja säätelyä nanomittakaavassa. CD44:n tapauksessa laajennamme nykyistä käsitystä sen ja hyaluronihapon vuorovaikutuksesta. Näytämme myös, kuinka CD44:n glykosylaatio voi moduloida hyaluronihapon sitoutumista muuttamalla sen sitoutumispaikkaa. JAK2:n tapauksessa osoitamme, kuinka sekä dimerisaatio että sitoutuminen solukalvoon ohjaavat signalointikompleksin aktivaatiota.

Työmme avulla saavutettiin uutta tietoa tutkittujen proteiinien toiminnasta ja vuorovaikutuksista. Tuloksista voi olla hyötyä lääkekehityksessä — erityisesti etsittäessä uusia lääkeaineiden sitoutumiskohtia esimerkiksi proteiinien glykosylaatio-, dimerisaatio- tai kalvoonsitoutumisrajapinnoista. Työ korostaa myös tietokonesimulaatioiden ja kokeellisten tekniikoiden yhdistämisen avulla saavutettua lisäarvoa tieteellisessä tutkimuksessa.

Preface

I started these doctoral studies at the Tampere University of Technology but quickly moved to the University of Helsinki, as our research group found a new home there. Since then, there has been several successes and failures along the way — and, I am grateful for all of them. It has been a privilege to work in such a stimulating environment, tackling interesting research questions with extremely talented people.

Firstly, I wish to thank Ilpo for allowing me to carry out these exciting projects in his research group. Without his knowledge and resources, this Thesis could not have been completed. I also would like to acknowledge the Finnish IT Centre for Scientific Computing (CSC) for providing the computer resources to carry out this work.

Next, I wish to thank my some of my closest colleagues. I am especially grateful to Hector for teaching me a lot about biomolecular simulations. Likewise, big thanks go to both Chetan and Vivek for doing excellent work with me on the JAK project. I would also like to thank Pavel for letting me visit his research group in Prague multiple times.

I am very thankful for all the current and former members of the group, both in Tampere and Helsinki. It has been a pleasure to travel with you to conferences and engage in many fruitful scientific discussions. Meetings outside the work have also been great fun. I am especially grateful to Giray, Maria, Misha, and Waldek for the epic scientific discussions we had during the lockdown periods of the COVID-19 pandemic.

While in Helsinki, friends outside of work have also been crucial in helping me settle into a new city and take my mind off science. Here, I wish to thank the great guys of GP laser tag, with whom I have had the pleasure to travel the world and cause mayhem.

Finally, I wish to thank my family for all the support.

Helsinki, May 2021
Joni Vuorio

Contents

List of Publications

I **Joni Vuorio, Ilpo Vattulainen, and Hector Martinez-Seara**. "Atomistic fingerprint of hyaluronan–CD44 binding". *PLoS Computational Biology*, 13(7): e1005663. 2017.

II **Joni Vuorio, Jana Škerlová, Milan Fábry, Václav Veverka, Ilpo Vattulainen, Pavlína Řezáčová, and Hector Martinez-Seara**. "N-Glycosylation can selectively block or foster different receptor–ligand binding modes". *Scientific Reports*, 11(5239). 2021.

III **Stephan Wilmes, Maximillian Hafer, Joni Vuorio, Julie A. Tucker, Hauke Winkelmann, Sara Löchte, Tess A. Stanly, Katiuska D. Pulgar Prieto, Chetan Poojari, Vivek Sharma, Christian P. Richter, Rainer Kurre, Stevan R. Hubbard, K. Christopher Garcia, Ignacio Moraga, Ilpo Vattulainen, Ian S. Hitchcock, and Jacob Piehler**. "Mechanism of homodimeric cytokine receptor activation and dysregulation by oncogenic mutations". *Science*, 367(6478): 643–652. 2020.

Author's Contribution

I The author co-designed the research with Hector Martinez-Seara and Ilpo Vattulainen. During the project, the author was solely responsible for preparing the simulation systems and performing the simulations. He analyzed and interpreted the data together with Hector Martinez-Seara and was primarily responsible for comparing the data to the literature. Finally, the author wrote the manuscript with supervision and comments from Hector Martinez-Seara and Ilpo Vattulainen.

II Similar to *Publication I*, the author co-designed the research with Hector Martinez-Seara and Ilpo Vattulainen. The author was also responsible for preparing, performing, and analysing the simulations — *i.e.*, the *in silico* part of the research. Both the author and Hector Martinez-Seara were involved in the interpretation and conceptualization of the simulation data. Hector Martinez-Seara was responsible for communicating the simulation findings with the NMR experts involved in this project. The author wrote the manuscript with inputs from all the co-authors, especially Hector Martinez-Seara.

III The author managed and performed the *in silico* part of this interdiscplinary research project. He also co-coordinated the discussion between the computational and experimental teams together with Ilpo Vattulainen. In practice, the author designed and prepared the simulation systems, performed most of the simulations, analyzed the results, interpreted the findings, and compared the results to experimental data. Chetan Poojari, Vivek Sharma, and Ilpo Vattulainen assisted in designing the simulation systems and interpreting the results. Chetan Poojari also performed a minor part of the simulations. Finally, the author implemented the *in silico* results into the manuscript with the help of the co-authors, especially Ilpo Vattulainen.

Symbols & Abbreviations

AA	All-atom
ATP	Adenosine triphosphate
CG	Coarse-grained
CHO	Chinese hamster ovary
COSY	Homonuclear correlation spectroscopy
cryo-EM	Cryogenic electron microscopy
DNA	Deoxyribonucleic acid
EC	Extracellular
ECD	Extracellular domain
EM	Electron microscopy
EpoR	Erythropoietin receptor
ER	Endoplasmic reticulum
FDA	U.S. Food and Drug Administration
FERM	Four-point-one, ezrin, radixin, moesin
FRAP	Fluorescence recovery after photobleaching
GAG	Glycosaminoglycan
GlcNAc	*N*-acetylglucosamine
GHR	Growth hormone receptor
HA	Hyaluronic acid, hyaluronan
HABD	Hyaluronic acid binding domain
HABP	Hyaluronic acid binding protein
HGR	Human growth hormone receptor
HSQC	Heteronuclear single quantum coherence
IC	Intracellular
JAK	Janus kinase
LINCS	Linear constraint solver
LJ	Lennard-Jones
MD	Molecular dynamics
MM	Molecular mechanics
mRNA	Messenger sibonucleic acid

NMR	Nuclear magnetic resonance
NpT	Isothermal–isobaric ensemble
NOESY	Nuclear Overhauser effect spectroscopy
NVE	Microcanonical ensemble
NVT	Canonical ensemble
PBC	Periodic boundary conditions
PBSA	Poisson–Boltzmann surface area
PD	Partially disordered
PDB	Protein Data Bank
PIP2	Phosphatidylinositol 4,5-bisphosphate
PK	Pseudokinase
PM	Plasma membrane or cell membrane
PME	Particle mesh Ewald
POPC	1-palmitoyl-2-oleoylphosphatidylcholine
POPS	1-palmitoyl-2-oleoylphosphatidylserine
PTM	Post-translational modification
QM	Quantum-mechanical
RMSD	Root-mean-square deviation
RNA	Ribonucleic acid
SASA	Solvent accessible surface area
STAT	Signal transducer and activator of transcription
TIRFM	Total internal reflection fluorescence microscope
TpoR	Thrombopoietin receptor
TK	Tyrosine kinase
TM	Transmembrane
TOCSY	Total correlated spectroscopy
tRNA	Transfer sibonucleic acid
UA	United-atom
VMD	Visual Molecular Dynamics
WHAM	Weighted histogram analysis method
ϵ_{ij}	Depth of Lennard-Jones potential well
ε_0	Permittivity of the vacuum
ε_r	Relative permittivity of the medium
ΔG	Free energy
ΔG_{bind}	Free energy of binding
$\Delta G_{\text{bind,vacuum}}$	Free energy of binding in vacuum
$\Delta G_{\text{solvation,complex}}$	Free energy of solvation for molecular complex
$\Delta G_{\text{solvation, subunit1,2}}$	Free energy of solvation for subunit 1 or 2
ΔG_{polar}	Free energy, polar component

$\Delta G_{\text{hydrophobic}}$	Free energy, hydrophobic component
$\Delta\Delta G_{\text{V617F}}$	Binding free energy difference between JAK2 wild-type and V617F
$\Delta\Delta G_{\text{V617F,exp}}$	Experimental binding free energy difference between JAK2 wild-type and V617F
ΔE_{MM}	Standard molecular dynamics energy terms
$\Delta S_{\text{normal mode}}$	Entropy estimated by a normal mode analysis
Δt	Time step
$\mathcal{H}_{\text{Total}}$	Total potential energy defined by force field
$\mathcal{H}_{\text{Bonded}}$	Potential energy of bonded interactions
$\mathcal{H}_{\text{Non-bonded}}$	Potential energy of non-bonded interactions
$\mathcal{H}_{\text{bond}}(r_{ij})$	Potential energy of bond stretching
$\mathcal{H}_{\text{angle}}(\theta_{ijk})$	Potential energy of angle
$\mathcal{H}_{\text{dihedral}}(\phi_{ijkl})$	Potential energy of dihedral twisting
$\mathcal{H}_{\text{LJ}}$	Potential energy of Lennard-Jones interaction
$\mathcal{H}_{\text{Coulomb}}$	Potential energy of Coulomb interaction
θ_{ijk}	Angle
θ_{ijk}^{eq}	Equilibrium angle
ϕ_{ijkl}	Dihedral angle
ϕ_{ijkl}^{0}	Reference dihedral angle
ξ	Distance-based reaction coordinate
ξ_{ijkl}	Improper dihedral angle
ξ_{eq}	Equilibrium improper dihedral angle
σ_{ij}	van der Waals distance
$\mathbf{F}$	Force
k_B	Boltzmann's constant
$k_{r,ij}$	Force constant of bond stretching
$k_{\theta,ijk}$	Force constant of angle bending
$k_{\phi,ijkl}$	Force constant of dihedral twisting
$k_{\xi,ijkl}$	Force constant of improper dihedral twisting
L	Simulation box length
m_i	Mass of a particle i
n	integer defining periodicity
q_i, q_j	Charges
r_{ij}	Bond length
r_{ij}^{eq}	Equilibrium bond length
$\mathbf{r}_i$	Position
t	Time
T	Temperature
$\mathbf{v}_i$	Velocity

Note: amino acids at specific positions in a protein are referred to with the single-letter code followed by their position in the canonical primary sequence of the polypeptide chain. For example, valine (V) at position 617 is written as V617. Amino acid substitutions at the protein level are indicated by adding the single-letter abbreviation of the substituting amino acid at the end of the code. For example, V617F denotes that amino acid 617 (valine, V) is mutated to phenylalanine (F).

Chapter 1

Introduction

Life consists of tiny functional units called cells [1]. They are microscopic systems evolved to adapt, propagate, and protect their hereditary information [2]. Different cells achieve these tasks with different means, with some operating as a single unit while others form complex multicellular organisms. Yet, every cell type shares a mutual chemical blueprint, governed by a few simple principles [3]. Firstly, they are enclosed in a membrane made of fat-like substances called lipids. Secondly, their hereditary information is encoded in a polymer of deoxyribonucleic acid (DNA). Finally, the information encoded in DNA is transcribed into proteins, which comprise the vast functional machinery that keeps the cells operating.

After construction and post-processing, proteins are localized in various intra and extracellular loci to perform their respective functions. For example, some of them are secreted from the cell to work as chemical messengers, such as cytokines, or structural units of connective tissues, such as collagen. Others remain inside the cell as structural units, such as actin. Furthermore, a significant fraction of all proteins end up in lipid membranes, where they perform critical signaling and transport functions [3]. The cell surface receptor proteins in the outer cell membrane or plasma membrane (PM) are especially important, as they receive cues from the immediate surroundings and subsequently convert them to intracellular signals that guide the actions of the cell. Due to their central role and location, they are also the prime targets of the pharmaceutical industry [4].

The operational principles of proteins are typically complex and involve interactions with several other molecules. For example, a large portion of eukaryotic proteins are decorated with carbohydrate moieties or glycans that share a vital — yet often unknown — role in the function of their parent protein [5,6]. Furthermore, in many cases, a particular protein needs

to be activated only under specific conditions, which are controlled by the binding of a messenger molecule called a ligand. If a disorder, such as a genetic mutation, forces the protein to be constitutively active or inactive even without its ligand, the resulting physiological imbalance can lead to the formation of a disease, such as cancer [3]. To treat such conditions, researchers try to understand how different proteins work, especially how they interact with other molecules and what causes them to malfunction in a pathological state. This field of uncovering the physical underpinnings of biomolecules' function is called molecular biophysics [2].

The backbone of biophysical research lies in various experimental methods. However, the small size of the studied biomolecules often poses a major challenge for these methods, as they are limited by the spatial and temporal resolutions required to probe such nano-sized objects. For instance, obtaining an Ångström-resolution view of a water-soluble protein typically requires crystallization of the target molecule, thus eliminating its dynamics. In many cases, the molecules under study are also so fragile that the measurement itself imposes a perturbation to their natural structure [7]. Moreover, it is challenging to visualize proteins in an *in vivo* environment to detect their full range of interactions with other biomolecules.

Fortunately, theoretical and computational methods can complement experiments by overcoming some of their typical limitations. Indeed, modeling and simulation techniques, such as molecular dynamics (MD) simulations, provide a particularly appealing route to study the interactions of proteins and other biomolecules with atomistic resolution. This has been made possible by improvements in both time and length scales reachable by simulations through novel multiscale approaches and ever-increasing computer capacity [8]. The accuracy of computational models is also constantly evolving. Especially when combined with appropriate experiments, simulations can help to decipher the molecular underpinnings of various diseases or to uncover the operational principles of physiologically important biomolecules and thus improve our health.

1.1 Objectives and Scope of the Thesis

This Thesis covers MD simulation studies of two PM-associated proteins and is thereby divided into two parts. The first part focuses on a receptor protein called CD44 [9–11] and its interaction with its carbohydrate ligand, hyaluronic acid (HA) also known as hyaluronan [11, 12]. The CD44–hyaluronan interaction is key in mediating cellular motility and adhesion in both health and disease, such as the metastasis of cancer [13]. The

interaction is also heavily modulated by *N*-linked glycosylations on the hyaluronan-binding domain (HABD) of CD44 [14–17].

The second part focuses on a protein family called Janus kinases (JAKs). They bind specific cytokine receptors at the PM and transmit essential chemical signals related to cell differentiation, proliferation, and survival *via* phosphate-transferring reactions [18, 19]. Their name stems from the two-faced Roman god, Janus, as they possess two nearly-identical phosphate-transferring domains. Importantly, specific mutations in JAK2 have been implicated in the onset of blood cancers, such as leukaemia [20].

The simulation results presented in this Thesis are predominantly linked to experiments conducted in collaboration with our partner groups. Yet, the focus of this work lies on communicating the *in silico* parts of each research project.

The first objective of this Thesis is to provide an atomic-scale understanding of the dynamics associated with the binding of HA to its CD44 receptor. Currently, there is one crystallographic 3D structure describing the CD44 HABD–HA interaction [11]. However, multiple studies have listed several HA binding amino acid residues that seemingly contradict with this crystallographic view [21–23]. While some of these studies have proposed a large-scale conformation shift to account for the observed discrepancies [10,23–25], others have coined the existence of secondary binding modes [10]. To gain a better understanding of the CD44–HA binding, we used MD simulations to couple HABD with its HA ligand in a spontaneous manner. The central question was whether there are secondary HA binding modes, and if so, what is their biological role.

The second objective is to decipher the role of glycosylations in mediating the CD44–HA interaction. Current findings show a correlation between the *N*-glycosylation of CD44 and the measured HA binding levels [14]. A negatively charged group of glycans called sialic acids — often associated with cancerous phenotypes — are known to act as prime regulators of HA binding [17,26,27]. Some studies also indicate certain glycosylation sites on CD44 HABD to be more important in controlling ligand binding than others [16]. Because of such findings, previous simulation efforts to study the glycosylations of CD44 have modeled CD44 HABD only in a partially glycosylated state without the HA ligand [27]. To obtain a more comprehensive picture, we modeled a fully *N*-glycosylated CD44 HABD, expressing several realistic glycan profiles, together with the HA ligand. Continuing the work from our first study, the key question was: can the glycosylations control receptor affinity by acting as switches between different HA binding modes. We also aimed to uncover the role of the different glycan profiles

in the recognition of HA.

The third objective is related to the second part of the Thesis: understanding the dimerization of JAKs. JAKs are a four-member family of signaling proteins constantly bound to a cytokine receptor with a 1:1 stoichiometry. In a normal state, the ligand-induced dimerization of such receptor–JAK complexes controls their activation as signaling molecules [18,19]. In a pathological state, however, oncogenic gain-of-function mutations keep the complexes constitutively active through an unknown mechanism — even in the absence of the ligand [28]. One plausible explanation is that these mutations, particularly the V617F mutation in JAK2, enhance the dimerization affinity of the cytosolic phosphate-transferring domains [29, 30]. Our central idea was to test this hypothesis using MD simulations. Additional goals involved deciphering the dynamics and relative organization of the four-domain structure of JAK2 as well as evaluating its interactions with a model lipid bilayer.

This Thesis places the work done into a broader context and evaluates it critically based on the works of others. The analysis culminates into a state-of-the-art view of the fields of CD44 and JAK research.

1.2 Structure of the Thesis

This Thesis is organized as follows. After the Introduction, Chapter 2 presents an overview of central biological concepts discussed in this Thesis. Here, the main focus lies on the two primary target proteins of this Thesis — CD44 and JAK2. The following Chapter 3 presents the theoretical background behind the methods used in this Thesis, focusing in particular on the MD simulation method. The Chapter closes by describing the simulation systems and models used in the included publications.

Chapters 4–6 present the findings of this Thesis. Chapter 4, based on *Publication I*, classifies three binding modes for the non-glycosylated CD44–HA interaction. Chapter 5 extends this work by exploring how glycosylations modify the receptor–ligand interaction. It is based on *Publication II*. Chapter 6 describes the dimerization of JAK2–cytokine receptor complexes in both healthy and pathological states. It is based on *Publication III*. Finally, Chapter 7 summarizes the findings and implications of this Thesis and discusses the potential future directions of the field.

Chapter 2

Biological Background

Living organisms are comprised of biomolecules. This Chapter begins by briefly introducing the fundamental concepts of three classes of biomolecules central to this work: lipids, proteins, and carbohydrates (Figure 2.1A–C). It then continues by showing how they combine in the environment of the PM. Nucleic acids, such as DNA (Figure 2.1D), form the fourth main class of biomolecules but are not the core topic of this Thesis, so they will not be discussed in detail. Similarly, worth noting are water and other small inorganic molecules that serve as the medium for all biochemical reactions but are not as such relevant for this Thesis. The Chapter closes by introducing the molecules of interest in this Thesis: CD44 and its HA ligand as well as JAKs and their cytokine receptor partners.

2.1 Membranes & Lipids

Our bodies are basically lumps of membranes. If we approximate that a typical human cell has a diameter of 10 μm, it then has a surface area of $4 \times \pi \times (10 \times 10^{-6}\,\mathrm{m})^2 = 12.6 \times 10^{-10}\,\mathrm{m}^2 \approx 10 \times 10^{-10}\,\mathrm{m}^2$. If we then consider that human has on average 40 trillion cells [32], the total surface area of these cells becomes $(40 \times 10^{12}) \times (10 \times 10^{-10}\,\mathrm{m}^2) = 40000\,\mathrm{m}^2$. To put this into perspective, this would equal to an area of 200×200 m^2. Moreover, this estimate is highly conservative, as it only accounts for the outer cell membranes even though there are more membrane compartments inside every cell. As a comparison, the typical surface area of human lungs is 50 to 75 m^2 while that of the skin is 1.5–2 m^2. It is therefore obvious that we are much bigger on the inside than outside — almost every physiological process is somehow connected to biological surfaces or membranes [1].

The PM is a barrier that separates the interior of a cell from the ex-

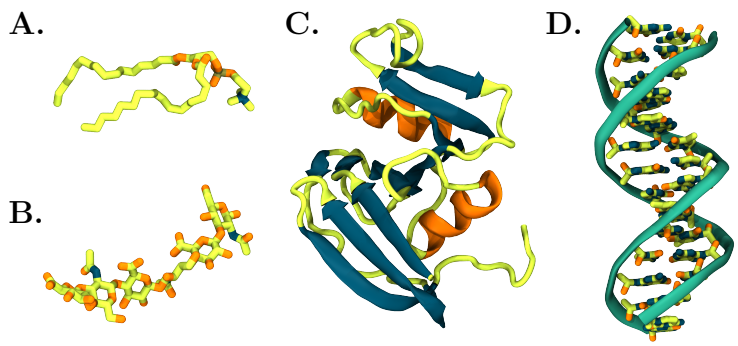

Figure 2.1: *Examples of four types of biomolecules. A) Example of a lipid, 1-palmitoyl-2-oleoylphosphatidylcholine (POPC). B) Example of a carbohydrate, hyaluronic acid. C) Example of a protein (single domain), CD44. D) Example of a stretch of double-stranded DNA. Images are generated with the Visual Molecular Dynamics (VMD) software [31].*

tracellular space [1]. It is an active boundary that controls the boarders, maintains vital concentration gradients, transports substances, interacts with the surroundings, and provides a basis for numerous biochemical reactions. Inside a cell, similar membranes define all the cellular organelles, such as Golgi apparatus, endoplasmic reticulum (ER), mitochondria, as well as all the other intracellular compartments evolved to perform their respective functions [1].

The many functions of the PM rise from its complex, mosaic-like structure [1]. It is a thin, tough, and hydrophobic film of lipids and proteins, decorated with various levels of attached carbohydrates located mainly on the extracellular side of the membrane. On the cytosolic side, the PM is tightly connected to the vast cytoskeleton network. From a physics perspective, the PM — like any biological membrane — is a dynamic, two-dimensional fluid, where most of the constituent molecules move laterally in the plane of the membrane, whose thickness is roughly 5 nm. The molecules are held together by noncovalent interactions, rendering the structure flexible and allowing for large-scale movements, such as cell division, without compromising the integrity of the barrier [1].

The basis for any biological membrane is the lipid bilayer [1,33]. Lipids are amphiphilic molecules, meaning that they possess both hydrophilic and hydrophobic moieties. The hydrophilic part can be as simple as a single hydroxyl group or a more elaborate polar region, such as a carbohydrate or alcohol moiety. The hydrophobic part typically constitutes fatty acid chains of varying chemical composition. When exposed to water, they are unable to participate in the hydrogen bonding network of the surrounding water molecules. Hence, to maximize the number of hydrogen bonds, the

surrounding water molecules find their minimum energy configuration by aligning around the hydrophobic parts of the lipids. Such alignment, however, increases the order and decreases the entropy of the water molecules compared to bulk water. Therefore, to reach a more energetically favorable state, water tends to interact with the hydrophilic regions of the lipids instead, such that the hydrophobic regions are forced together and away from the water molecules. This so-called hydrophobic effect gives rise to various lipid structures — most notably the bilayer [1, 34].

Lipids are a broad class of molecules. Fatty acids and their numerous molecular derivatives can all be generally categorized as lipids. These molecules have various structures, whose common feature is their partial insolubility in water. Moreover, their functions are as diverse as their structures. In the human body, fats and oils are the main forms of energy storage, phospholipids and sterols serve as structural elements of biological membranes, while other less-abundant lipid classes share crucial roles as hormones, hydrophobic anchors, cofactors, electron carriers, messengers, or emulsifying agents. It is estimated that biological membranes contain hundreds of thousands of different lipid species'. Yet, only the most abundant of these membrane lipids, glycerophospholipids, bear relevance for this Thesis. For a more thorough description of all the lipid types, see Refs. [1, 3, 33].

Glycerophopholipids are the most abundant class of lipids in biological membranes [33]. Although several lipid species belong to this class, they all share a common structural architecture. Namely, the backbone of these molecules is formed by a glycerol moiety. The hydrophobic regions comprise two fatty acid chains esterified to the first and second carbon of this glycerol backbone. These fatty acids typically have 12–24 carbon atoms which can be either saturated — *i.e.*, there are no double bonds between the carbons — or unsaturated with one or more double bonds. The hydrophilic region is a polar alcohol head group, joined to the third hydroxyl group of the glycerol backbone through a phosphodiester bond. The greatest variance between the various glycerophospholipid species' originates from their differences in the polar head group, as there is a plethora of possible compounds that can be chemically linked to the glycerol backbone [33].

The names of glycerophospholipids are derived from their parent compound, phosphatic acid, according to the type of the fatty acid tails and the polar head group. For example, name 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine or POPC signifies that palmitoyl and oleyl fatty acids are esterified to the first and second hydroxyl group of the glycerol backbone, respectively, while a choline alcohol is esterified to the third hydroxyl position [33].

2.2 Proteins

Almost all biomolecules in our bodies are either proteins or products of their action [3]. Proteins constitute the nanoscale machinery that executes the diverse array of physiological functions our bodies require to stay alive at every moment. These functions include acting as signaling agents [35, 36], transporting substances over membranes [37], maintaining the structure of our tissues [38], generating movement [39], establishing concentration gradients [40], initiating signaling [41], reading and duplicating DNA [42], as well as catalyzing the production of other biomolecules [43].

Proteins are formed from 20 amino acid types linked together in a linear chain *via* covalent peptide bonds [3]. The sequence of this polypeptide chain is dictated by the nucleotide sequence of its parent gene. Beyond this primary sequence, the biological activity of proteins originates from their higher levels of structural organization. That is, the linear polypeptide chains fold into different three-dimensional (3D) structures that ultimately determine their biological function. However, even completely disordered structures do not always lack function, as many proteins are known to operate in an unfolded state under physiological conditions. Additionally, a wide array of post-translational modifications (PTMs) [3], such as glycosylations, modify protein structures and thereby also their localization, role, and degradation [44, 45].

2.2.1 Structural Characterization of Proteins

The structure of a protein determines its function. There are four main levels in the structural organization of proteins: primary, secondary, tertiary, and quaternary [3]. Primary structure describes the sequence of amino acids in a polypeptide chain and thus sets the basis for the other, higher levels of organization. Similar amino acid sequences share uniform folding patterns, generating homologous proteins or homologs. Often two homologs are separated by only a few alterations in their amino acid sequences. Such single amino acid changes can be conservative, non-conservative, or invariant depending on the chemical similarity between the exchanging amino acids.

Secondary structure describes the three dimensional folding of local segments of the polypeptide chain [3]. There are three main categories of secondary structures in proteins: α helix, β strand, and turn. They each have well-defined features, such as turn length and translational distance, which can be used to characterize these structures. Secondary structures are held together by precise hydrogen bonding networks between subsequent amino

acids in the chain.

Tertiary structure constitutes the overall topology of the polypeptide [3]. For small (< 150 residues) soluble proteins, this topology predominantly entails a globular unit, which can also be referred to as the overall fold of the polypeptide. Large proteins (> 150 residues) may be organized into multiple structural units called domains. These autonomously folding units have formed due to a gene duplication or fusion and can often perform their respective functions without the rest of the protein. The overall fold of a domain can be expressed based on its predominant secondary structure elements. For example, domains can be described as α, β, or mixed α/β folds. Tertiary structure is maintained by both chemical (disulphide bridges) and physical (hydrogen bonds, van der Waals forces, hydrophobic effect, and electrostatic interaction) interactions.

Quaternary structure involves more than one polypeptide chain. Here, the individual chains are termed subunits [3]. Proteins usually form active complexes with each other, allowing for allosteric regulation and formation of novel binding sites for their ligands into the areas between the subunits. These protein complexes are held together by non-covalent, physical interactions between the constituent molecules.

2.2.2 Protein Synthesis, Quality Control & Mutations

The flow of genetic information from DNA to RNA to proteins occurs in two steps: transcription and translation [3]. Transcription refers to the synthesis of a single-stranded messenger RNA (mRNA) copy of a segment of two-stranded DNA. This synthesis is performed by a protein called RNA polymerase and facilitated by several transcription factors. After additional processing and splicing, the mRNA strand is transported from the nucleus to the ER. The second step in the flow of genetic information, translation, refers to the synthesis of proteins from the mRNA template, occurring in the ribosomes of the ER. Ribosomes read the mRNA template as units of three nucleotides or codons at a time. Special start and stop codons define the reading frame by designating where to start and stop. Translation begins when soluble transfer RNAs (tRNA) bind into the ribosome–mRNA complex so that the codons from the mRNA and tRNA match. Each tRNA unit confers one amino acid that the ribosome joins to the growing polypeptide chain through a polypeptide bond. The protein is ready after folding and possible attachment of PTMs in the ER and Golgi complex.

The four-nucleotide code of DNA can produce $4^3 = 64$ different three-nucleotide combinations. This would allow for 64 different amino acids if one excludes the designated start and stop codons. However, we only

have 20 different amino acids, meaning that several codons code for the same amino acid. This genetic code is well known and applies to all living organisms [1].

DNA replication is the process by which DNA copies itself during cell division [3]. Due to the proofreading capabilities of DNA polymerase — the protein responsible for this duplication — the error rates are as low as 1 in every 100 million bases [46]. However, when these improbable but inevitable errors occur in the protein-coding segments or exons of a gene, they can affect the structure and function of the proteins produced. Several chemical and physical factors, such as radiation [47], can also induce mutations. If the effect of a genetic mutation reaches the protein level it can lead to alterations in the phenotype of the organism. Such alterations are often neutral or harmful to their carriers, with examples ranging from color blindness [48] or lactose intolerance [49] to cancer [50] or death.

Small-scale mutations involve an alteration of one or a few nucleotides within the genome [3]. As a prime example of a small-scale mutation, point mutations change a single nucleotide in the sequence. Depending on the change, they might cause a switch of a single amino acid in the final protein product. If the chemical nature of an amino acid is changed, for example from basic to acidic, the behavior of the protein might change in a clinically relevant manner. Point mutations can also introduce a premature stop codon, resulting in a truncated protein product. Frameshift mutations, on the other hand, disrupt the DNA sequence by non-multiples of three nucleotides. They are caused by insertion or deletions of nucleotides to or from the sequence. Such a complete change in the reading frame typically leads to a severely altered end product, which can be very detrimental to the organism carrying the mutation [1].

2.2.3 Membrane Proteins

Biological membranes are crowded with proteins. Indeed, these membrane proteins constitute roughly 50 % of the volume of the PM [51] and approximately 30 % of the entire human proteome [52–54]. While the structural roles of the PM are handled primarily by lipids, proteins carry out most of its other functions, including the transduction of extracellular signals through the membrane into the cytosol *via* receptor proteins, such as ligand-gated ion-channels [55], enzyme-linked receptors [41, 56], or G protein-coupled receptors [57]. These proteins bind their respective ligands on the extracellular side of the PM, resulting in a conformational change of one protein or dimerization of two or more protein subunits. Such structural changes then activate the receptors by opening a channel for ions, ini-

tiating cytosolic enzyme function, or inducing the binding of intracellular activator proteins. This activation, in turn, initiates chemical signaling cascades within the cell that result in the regulation of gene transcription [1]. Due to the physiological importance of cellular signaling, membrane receptors serve as targets for ca. 50 % of modern pharmaceuticals [4].

Membrane proteins are divided into two main categories — integral and peripheral membrane proteins — according to their mode of interaction with the membrane [1]. Integral membrane proteins are permanently anchored to the hydrophobic lipid core. This large group of proteins is further divided into transmembrane proteins, whose polypeptide chain spans the lipid bilayer either one or multiple times, and integral monotopic proteins that attach covalently to only one leaflet of the bilayer through a lipid anchor. The membrane-spanning segments of transmembrane proteins typically involve a varying number of α helices composed of mainly hydrophobic amino acid residues. As an example, enzyme-linked receptors, such as receptor tyrosine kinases (RTKs) [41], possess a single membrane-spanning helix, while G protein-coupled receptors, such as glucagon receptor, cross the membrane with seven transmembrane helices [57].

Peripheral membrane proteins are temporarily attached to the membrane lipids or proteins, *e.g.*, through electrostatic interactions and lack the membrane-spanning component [1]. They therefore readily dissociate if a polar reagent, such as a concentrated salt solution, is applied to the membrane.

2.2.4 Enzymes

Enzymes constitute a significant fraction of all proteins. They are a vastly diverse group of biological catalysts that work by accelerating biochemical reactions [3]. Many membrane proteins are also simultaneously enzymes, catalyzing a particular signaling reaction on the cytoplasmic side of the PM in an activation-dependent manner.

Given the vast heterogeneity of enzymes, their names are usually derived based on their substrates or the chemical reaction they catalyze, with the word ending in *-ase* [3]. They are categorized into six primary classes, such that oxidoreductases catalyze oxidation/reduction reactions; transferases transfer functional groups; hydrolases catalyze the hydrolysis of chemical bonds; lyases cleave bonds by other means than hydrolysis; isomerases catalyze the isomerization of a molecule; and ligases join molecules together with covalent bonds [3]. All these classes are further divided into multiple lower-level categories based on their specific roles. For example, kinases are transferases that catalyze the chemical transfer of a phosphate group.

2.3 Carbohydrates

Carbohydrates or polysaccharides act as storage fuels, information carriers, and structural modulators of other molecules [3]. On the surface of the PM, they form the glycocalyx layer that protrudes outwards from the membrane (Figure 2.2). This layer provides recognition sites for extracellular molecules and pathogens, thus enabling communication between cells and their surroundings. The polysaccharide chains in the glycocalyx are typically connected to membrane proteins or lipids.

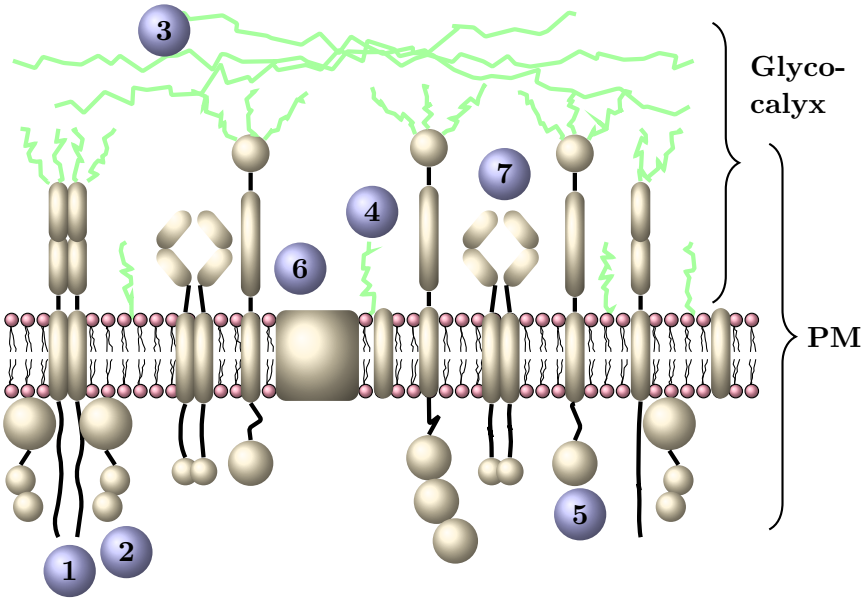

Figure 2.2: *Schematic figure of the glycocalyx and plasma membrane. Color coding is as follows: pale depicts proteins, green describes carbohydrates, and red beads are lipid headgroups. Blue spheres with numbers highlight key features of the PM environment. Glycoprotein dimer (1) spans the membrane and is bound to peripheral membrane proteins (2) on the cytosolic side. The glycocalyx is formed from long, linear glycosaminoglycans (3), glycolipids (4), and glycoproteins on the extracellular side of the PM. Channel proteins control the passage of substances through the lipid membrane (6), while signaling receptors (7) convey the signals from the extracellular space into the nucles.*

Glycocalyx Expands the Plasma Membrane

Glycocalyx is a carbohydrate-rich layer lining the PM of almost all cell types from simple bacteria to complex Eukaryota [58]. It is formed primarily from polysaccharides, which are either soluble or conjugated to other

biomolecules (Figure 2.2). The size and shape of the glycocalyx vary in a cell type-dependent manner. Bacterial cells can have only a few cell-surface oligosaccharide chains, while specific eukaryotic cells can express macroscopic carbohydrate-rich structures, such as the hair-like appendices on the blood vessel endothelium [58–61]. The primary functions of glycocalyx include the gating of substances by formation of a functional interface between cells and their surroundings [58].

Glycosaminoglycans (GAGs) encompass a class of biopolysaccharides unique to animal cells. They constitute the bulk of the carbohydrate content in the most prominent eukaryotic glycocali, such as the ones found in stromal or endothelial cells [58]. Unlike common polysaccharides like starch or glycogen that are composed of a single repeating monomer, GAGs are polymers of repeating disaccharide units. This repeating disaccharide unit consists of uronic acid, such as D-glucuronic acid or L-iduronic acid, and hexosamine, such as *N*-acetylglucosamine or *N*-acetylgalactosamine [58]. Additionally, their varying sulfation patterns often give these structures an added level of complexity. Due to the carboxylate group in the uronic acid, GAGs possess a negative net charge and, as a result, assume extended conformations in solution [62]. GAGs conjugated to proteins and lipids form proteoglycans and lipopolysaccharides, respectively. These conjugate structures bind different extracellular ligands and act as structural units, maintaining the strength and resilience of cells and tissues [58].

N-linked Glycans in Protein Structure and Function

Glycoproteins are another protein-linked class of glycoconjugates. Similar to proteoglycans, they also house carbohydrates of varying complexity attached to a core protein. However, glycoproteins generally have shorter (1–20 monomers), more branched, and more diverse carbohydrate residues attached *via* specific glycosidic linkages [58]. On a broader perspective, glycosylation is one of the most common PTMs. It is estimated that over 50 % of all of the mammalian proteins are glycosylated and 1 % of all genes are coding for enzymes related to glycosylation [3]. The attached glycans shield the parent protein from proteolytic cleavage, stabilize its specific conformations, participate in recognition events, and serve as specific name tags for transport [63].

The linking of oligosaccharides occurs primarily with two mechanisms: *O*-linked and *N*-linked glycosylation [5]. In the *O*-linked glycosylation, an oligosaccharide attaches to the hydroxyl group of a serine or threonine residue *via* a so-called *O*-glycosidic link. In the *N*-linked glycosylation, an oligosaccharide attaches to the amide nitrogen of an asparagine residue *via*

a so-called N-glycosidic link. The N-glycosidic linking requires a consensus sequence of three amino acids, where the first is the linking asparagine, the second is anything else but proline, and the third is either serine or threonine [3]. In such sequence, the hydroxyl group of the serine/threonine residue forms a hydrogen bond with the carbonyl group of the asparagine side-chain. This, in turn, increases the nucleophilicity of the amide group in the asparagine side-chain, thus increasing its affinity for the saccharide. Yet, the overall propensity of an asparagine residue to be glycosylated depends also on other factors than its position in the primary sequence, such as the tertiary structure around the consensus sequence.

Biosynthesis of N-glycans occurs in stages [5]. In the first stage, specific enzymes assemble a precursor glycan by a stepwise transfer of individual glycan units to a lipid carrier (dolichol pyrophosphate) on the cytoplasmic face of the ER. At a certain point, flippase enzymes recognize the structure of dolichol pyrophosphate with two attached N-acetylglucosamine (GlcNAc) sugars and five mannoses (Man) and translocate it to the lumenal surface of the ER membrane. On the lumenal surface, further mannose residues are attached to the lipid-linked precursor until the final glycan, $\mathrm{Glc_3Man_9GlcNAc_2}$, is recognized by an oligosaccharyltransferase that catalyzes its attachment to an asparagine residue of a newly-synthesized polypeptide chain. In the second stage, this glycan undergoes trimming and processing in the ER through conserved glycan remodeling pathways. At this stage, misfolded glycoproteins are eliminated and correctly folded proteins are further trimmed to produce the final N-linked glycans [5].

After the biosynthesis, all N-glycans share the same $\mathrm{Man_3GlcNAc_2}$ core structure, referred to as the core pentasaccharide [5]. The extensions of this core — *i.e.*, antennae — then define the N-glycans into one of three general classes: high mannose, complex type, or hybrid [5]. High mannose forms contain only mannose in their antennae. Complex type N-glycans share a general $\mathrm{Neu5Ac}_X\mathrm{Man_1GlcNAc_1}$ structure in all their antennae, where Neu5Ac refers to N-acetylneuramic acid (sialic acid) and X usually equals 1–3 [64]. Complex type N-glycans also have a fucose residue attached to the first GlcNAc linked to the asparagine. Hybrid types are a combination of the other two classes with some antennae housing high mannose and others having complex type-specific saccharides. Despite this classification, N-glycans — and glycans in general — typically experience high structural variance [5,63]. Incorrect carbohydrate structures are generally not corrected by the cell as effectively as, *e.g.*, misfolded proteins [3]. Consequently, N-glycan profiles of protein extracts typically entail large populations of different glycan structures [65].

Hyaluronic Acid

Hyaluronic acid or hyaluronan (HA) is a natural glycocalyx-related GAG polymer, which — unlike other GAGs — is not conjugated to proteins or lipids nor sulfonated. Its linear and repeating structure consists of *N*-acetyl-D-glucosamine (GlcNAc) and D-glucuronic acid (GlcUA) residues connected *via* a $\beta(1 \rightarrow 4)$-glycosidic linkage [12]. These disaccharide units are, in turn, linked *via* $\beta(1 \rightarrow 3)$ bonds to form polymeric HA. In normal physiological conditions, HA polymers can consist of ca. 2000–2500 disaccharide units, spanning lengths of 2-25 μm and molecular weights of up to 6-7 $\times 10^6$ Da [12, 66].

Like other GAGs, HA polymer is polyanionic due to the carboxyl groups of GlcUA. Yet, HA also has a hydrophobic face due to the axial hydrogens falling into one face of the disaccharide. This structure gives HA unique properties, such as an internal swelling pressure, which allows it to act like a hydrogel — an entangled polymer network that retains water in its structure. Hence, HA serves as a lubricant and gives tissues, such as synovial fluid and vitreous humour, their jelly-like consistency [12].

Besides its structural roles, HA mediates a variety of physiological functions, such as embryonic development, immune response, wound healing, cell motility, and angiogenesis [12]. It also plays a pivotal role in the pathogenesis of various disease, like diabetes, cancer, pneumonia, kidney diseases, and asthma [12]. This high diversity in HA-related physiological functions originates from a large number of hyaluronan-binding proteins (HABPs) [67] that mediate the many functions of HA with different tissue-specificities, cellular localizations, and regulation mechanisms.

A striking example of the effect of HA on the body is apparent in the naked mole-rat. This mouse-sized rodent expresses abnormally high molecular weight HA polymers in its tissues, rendering it immune to cancer and enabling it to have a life expectancy of over 30 years [68]. As a comparison, a mouse of similar size would have a maximum life expectancy of two to four years. Such findings underline the importance of understanding the molecular basis of HA recognition by proteins and how it is regulated.

2.4 CD44 Glycoprotein Binds Its Hyaluronan Ligand

CD44 is a type I transmembrane protein that, as its primary function, binds HA [9]. The binding controls cellular adhesion and HA-related signaling through cytosolic accessory proteins [69]. CD44 is encoded by a

single gene that can produce multiple isoforms, due to the alternative splicing of exons encoding a segment of the juxtamembrane stalk domain. The canonical form of CD44 consists of 723 amino acid residues that divide into four domains: hyaluronan binding domain (HABD), stalk domain, transmembrane domain, and cytosolic region (Figure 2.3A). From these, only HABD has been structurally resolved. This 158-residue globular domain is unaffected by splicing and performs its task of binding the HA ligand even in the absence of the rest of the protein. It consists of a conserved α/β fold (Figure 2.3B), called the Link module [10], which is also present in other carbohydrate-binding proteins, such as TSG-6 [70] and LYVE-1 [71]. In CD44 HABD, the canonical Link module is further extended by N and C-terminal flanking regions, and stabilized by three disulphide bridges, together forming the structurally rigid HA binding unit [10].

Binding of CD44 HABD to HA oligomer has been characterized with X-ray crystallography [11]. This structure illustrates how HA attaches in a key-to-lock manner to a shallow binding groove on the Link module (Figure 2.3C–D). One of the most striking features of this crystallographic binding includes a well-matching surface topology, with a hydrophobic pocket able to accommodate the methyl group of a bound GlcNAc residue. Another noticeable feature is the switch-like behaviour of the nearby R41 residue that can assume two different conformations termed as A and B-forms. The A–B switch is initiated by torsion in the backbone of the $\beta1$–$\alpha1$ loop (Figure 2.3B) that causes the R41 to flip ca. 90° [72]. While the binding of the HA ligand is the same in both R41 conformations, the B-form shares a more intimate interaction with the bound ligand [11]. In this interaction, the guanidino group of R41 side-chain renders charged interaction with the hydroxyl group of the bound HA (Figure 2.3D). Hence, the B-form has been thought to represent a high-affinity conformation stabilized by the binding of HA [11]. Moreover, in previous mutagenesis studies, R41 has been pinpointed as a critical HA binding residue, as its substitution to alanine completely terminates the CD44–HA interaction [21, 22].

A second, more pronounced conformation shift occurs in the C-terminal extension of HABD and also influences the ligand binding [23, 24]. This shift from ordered (O) to partially-disordered (PD) conformation involves the partial unfolding of the C-terminal flanking region (*i.e.*, strands $\beta8$ and $\beta9$) [24]. As a result, the increased conformational freedom enables multiple positively-charged amino acid residues at the flexible C-terminus to interact with the bound ligand [73]. Studies indicate that HABD interconverts between the O and PD forms spontaneously in solution, but the PD conformation is more favorable in the ligand-bound form of the protein [24].

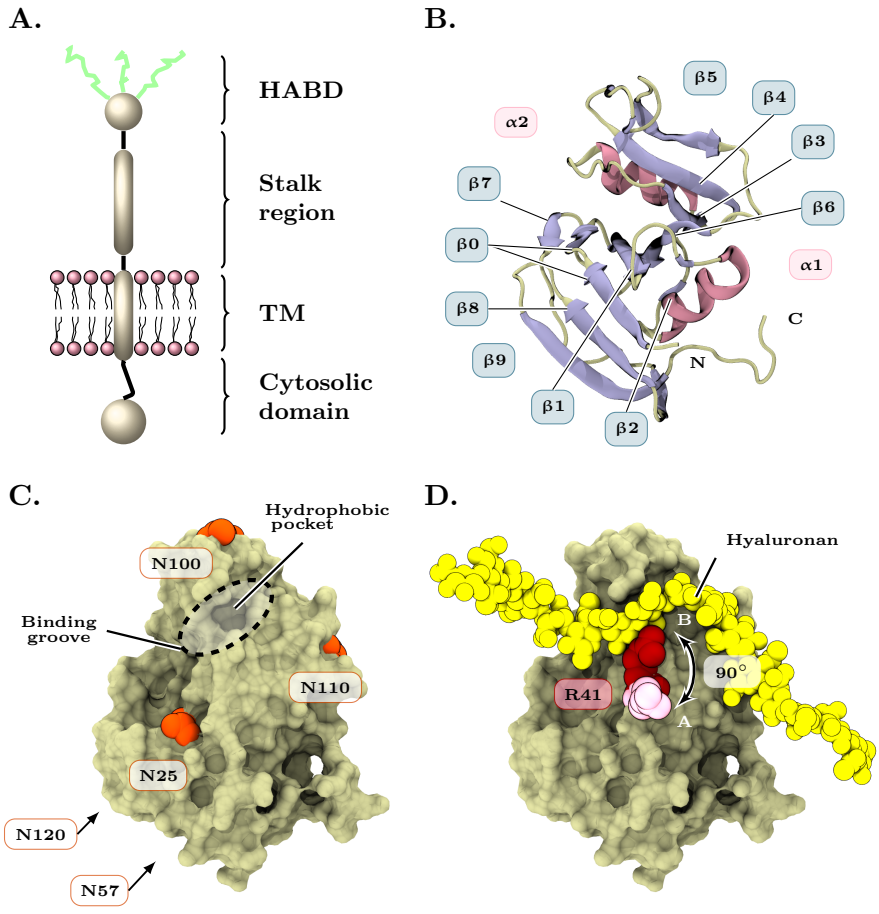

Figure 2.3: *CD44 and its hyaluronic acid-binding domain (HABD). A) Schematic presentation of the full structure of CD44. B) Naming of the secondary structure elements on HABD. The structure is extracted from PDB:1UUH [10]. C) HA-binding (dashed circle) and N-glycosylation (orange beads) sites on CD44 HABD. Residues N120 and N57 fall on the other side of the domain, as indicated by the arrows. The orientation of the protein in this surface representation is identical to the cartoon representation in B. D) Crystallographic HA (yellow beads) binding on HABD [11]. The most important HA binding residue, R41, is depicted with red in both of its possible conformations.*

The binding of HA to CD44 is also heavily regulated by the *N*-glycosylation of HABD [14–17, 74, 75]. It contains five *N*-glycosylation sites (N25, N57, N100, N110, and N120) [10] from which three (N25, N100, and N110) are

juxtaposing the crystallographic HA binding groove. Site N25 locates in the middle of the HA binding face next to two key HA binding arginines R41 and R78. Sites N100 and N110 lie on the top portion of the HABD, as presented in Figure 2.3C. Sites N57 and N120, on the other hand, reside on the opposite side of HABD that lacks any known HA binding residues. Structurally, the *N*-glycosylation sites are known to be primarily occupied by triantennary complex type oligosaccharides, especially in cancerous cell types [16,65,75]. Such an *N*-glycan profile means that a significant portion of the molecular weight and charge of CD44 HABD originates from the saccharide components.

N-glycans have been shown to define three levels of CD44-dependent HA-binding [14]. Interestingly, a single cell line can be manipulated with glycosidases to express all of these three levels. The first level, observed in untreated cells, is completely inactive to HA binding. The second level is also inactive but can be induced to bind HA with specific antibodies or glycosidases. At the third level, the binding is constitutively active. Glycan-cleaving enzymes, such as tunicamycin and neuraminidases, drive CD44-positive cells towards the active HA binding level [14, 15, 75], implying that *N*-glycans on CD44 interfere with the binding of HA. However, some studies have also reported contrasting findings, observing tunicamycin treatment to suppress ligand binding in certain tumor-derived cell lines and reduce their ability to bind immobilized HA [76, 77]. Based on these findings, CD44 requires at least some level of glycosylation to bind HA.

Explaining the contrasting findings, the *N*-glycans of CD44 have been found to elicit a dual effect on HA binding [17]. First, GlcNAc residues at the root of the complex *N*-glycan core have been found to enhance the binding of HA *via* an unknown mechanism. Second, sialic acids cappings in the glycan antenna have been observed to inhibit the binding of HA [14–17, 26, 75, 78]. In fact, the three distinct levels of HA binding are thought to originate from a varying number of sialic acids, as glycans in the cells of the inducible HA binding phenotype are known to contain more α2,3-linked sialic acids than the constitutively active cells [17]. The inactive cells, on the other hand, are thought to express mainly α2,6-linked sialic acids, thus rendering them harder to cleave with typical α2,3-specific neuraminidases [16]. Such regulation by sialic acids can be intuitively explained by charge repulsion, as both HA and sialic acids share a negative charge.

2.5 Cytokine Signaling and Janus Kinases

Cytokines are protein messengers secreted by various cells into the bloodstream and extracellular space [3, 36]. They act much like hormones, and the distinction between these two classes of messenger molecules is often nominal. Typical cytokine-regulated physiological processes include the immunity response [79] and the production of blood cells and platelets [35].

Cytokines act through receptor proteins [28]. These receptors are a diverse group of transmembrane proteins that bind cytokines and regulate growth, survival, and maturation of their parent cells through enzymatic kinase activation. The efforts to classify this group of proteins is largely based on their structural features and principles of activation. Among the various cytokine receptor classes, class I and II receptors lack any intrinsic signaling capabilities and therefore rely on Janus kinases (JAKs) for the enzyme function [80] (Figure 2.4A). Structurally, cytokine receptors consist of an extracellular part containing the ligand binding site, a single-pass transmembrane (TM) helix, and an intracellular (IC) part that is presumed to be mostly disordered — *i.e.*, lacking a definite fold [36].

JAKs are a family of cytosolic peripheral membrane proteins. These non-receptor tyrosine kinases regulate central physiological processes, such as immunity, development, and growth, *via* a constitutive association to the intracellular parts of their cognate cytokine receptors [19]. The JAK family includes four members, JAK1, JAK2, JAK3, and TYK2, which are structurally homologous, yet activated through different cytokine receptors [56]. The activation occurs generally *via* the dimerization of two receptor–JAK complexes. This enables the *trans*-autophosphorylation of the dimerized JAKs and leads to the activation of the signaling complex [56]. The receptor–JAK complexes are typically heterodimers, meaning that both the receptor and JAK components are different between the two subunits. However, certain cytokine receptors, such as erythropoietin receptor (EpoR) and growth hormone receptor (GHR), form only homodimers and bind exclusively to JAK2 [28]. In these cases, it is unclear whether the receptors dimerize upon ligand binding or exist in a pre-dimerized form that undergoes some structural rearrangement upon ligand binding. That is, there is evidence supporting both claims [86, 87].

The activation of the receptor–JAK complex leads to the canonical intracellular JAK–STAT signaling pathway [19]. It begins when two adjacent JAKs *trans*-autophosphorylate each other *via* their protein kinase domains and subsequently transfer the phosphates to the intracellular part of the cytokine receptor, which leads to the recruitment and phosphorylation of transcription factors called STATs (signal transducer and activator of tran-

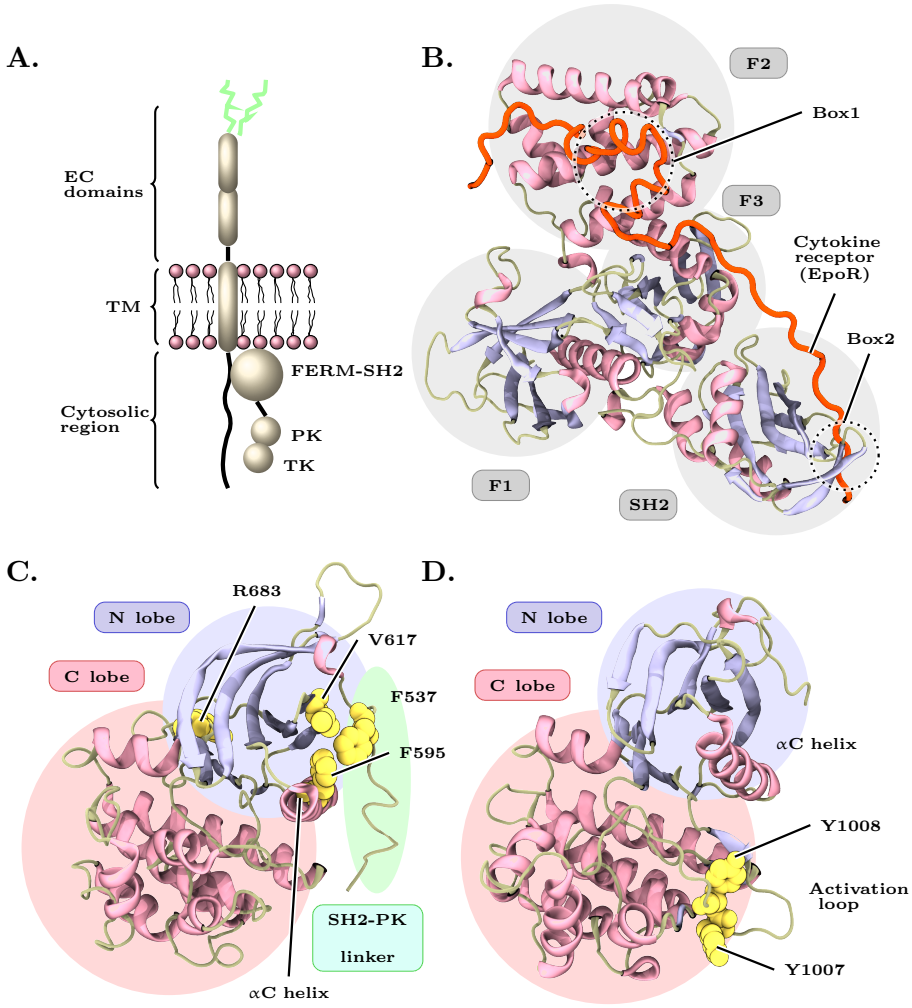

Figure 2.4: *JAK2 and its domains. A) Schematic presentation of a JAK2 monomer bound to a cytokine receptor. B) FERM-SH2 domain of JAK2. The subdomains of the FERM-SH2 structure are indicated with markings and gray circles (background). The intracellular region of erythropoietin receptor is shown in orange. The structure is extracted from PDB:6E2Q [81]. C) The pseudokinase (PK) domain of JAK2. The structure is modeled from PDB:4l00 [82] and PDB:5UT3 [83]. D) The tyrosine kinase (TK) domain of JAK2. The structure is modeled based on PDB:4OLI [84] and PDB:4IVA [85]. The most important structural elements are highlighted in C and D.*

scription). The final step in the cascade is the translocation of the activated STAT dimers to the nucleus where they regulate their target genes.

The structure of JAKs comprises ca. 1200 residues divided into four conserved domains that each have their role in the function of the protein. At the N-terminal end, the FERM (band 4.1 protein, ezrin, radixin, moesin) and SH2-like (Src homology-2) domains form a holodomain (Figure 2.4B), which controls the recognition of distinct cytokine receptors *via* two known binding sites for the membrane-proximal Box1 and Box2 sequences of the receptors [88–90]. The FERM domain is further divided into subdomains F1, F2, and F3 based on their evolutionary origin. Downstream in the sequence, JAKs have two canonical kinase domains, pseudokinase (PK) [91] and tyrosine kinase (TK) [92], from which the latter is responsible for the enzymatic activation of the protein (Figure 2.4C–D). The C-terminal TK domain is activated through a *trans*-autophosphorylation of two tandem tyrosines, Y1007 and Y1008, located in the activation loop (Figure 2.4D). The activated TK domain then phosphorylates specific tyrosines in both the cytokine receptor and the recruited STAT proteins. The preceding PK domain is structurally homologous to the TK domain but misses the kinase activity due to the lack of critical residues required to transfer the phosphate from ATP to the substrate [93]. Instead, it plays a role in the regulation of the TK domain by binding to its hinge region and stabilizing its inactive conformation. However, the full extent of this regulation mechanism is not known.

The prevailing paradigm states that the autoinhibitory interaction between the PK and TK domains prevents the *trans*-phosphorylation of the activation loop by stabilizing the inactive state of the TK domain [84]. This view stems from the X-ray structure of TYK2 PK–TK interaction as well as a similar molecular dynamics simulation-derived structure of JAK2 PK–TK interaction [84]. In this autoinhibitory interaction, the N lobe of the PK domain is connected to the N-lobe and the hinge region of the TK domain. As the activation loop of the TK domain is not directly located at this interface, this pose most probably acts by hindering the lobe movements of the TK domain crucial for its kinase activity. There are also two known negative-regulatory phosphorylation sites, S523 and Y570, on the PK domain. While the exact role of these residues is unclear, Y570 phosphorylation is increased upon cytokine activation, suggesting that the TK domain might phosphorylate it as a possible feedback mechanism [94].

Heterozygous gain-of-function mutations in the genes related to the JAK–STAT pathway cause a range of physiological disorders [20,95]. From these, the most common are myeloproliferative neoplasms (MPNs) — a family of bone marrow disorders, which serve as precursors for blood cancers, such as leukaemia [96]. The most common MPNs include essential

thrombocythemia, polycythemia vera, and primary myelofibrosis. They are caused predominantly by hyperactivating mutations located in the PK domains of JAKs [97–100]. These mutations force the JAK-mediated signaling to be constitutively active, which leads to an overproduction of blood cells/platelets in the cells of the bone marrow. Yet, the exact molecular mechanisms of these mutations remain elusive. The most commonly identified mutation is the V617F substitution in the PK domain of JAK2, which is responsible for over 95 % of polycythemia vera and roughly half of the cases of essential thrombocythemia and primary myelofibrosis [101]. Furthermore, an array of other oncogenic mutations are known to exist, especially in the SH2-PK linker segments of JAKs [30].

Unlike other activating mutations, such as R683G or L611S, V617F does not reside at the PK–TK autoinhibitory interface [84]. It is instead located proximal to the unstructured SH2-PK linker. Due to this localization, the mechanism of action for the V617F mutant has remained enigmatic [29]. It has been proposed that the mutation causes a conformational pressure to the nearby phenylalanine residues, exposing F537 which is normally covered by F595 of the αC helix [82]. However, the link between this putative conformational switch and hyperactivation of JAK2 is missing. V617F has also been proposed to destabilize the SH2–PK linker [102], but this has also remained as a speculation with little proof. Additionally, there are certain rescuing mutations, such as F595A, that can suppress the V617F activity [30]. However, F595A also suppresses other activating mutations distant from F595, indicating that F595A could act by partially destabilizing the structure of PK.

A plausible hypothesis states that V617F mediates the ligand-independent dimerization of the PK domains [29]. Such abnormal dimerization of JAK2 would be driven primarily by the cytoplasmic domains, thereby removing the need for external cytokine activation. The dimerization of the PK domains could potentially also disrupt the autoinhibitory PK–TK interaction, thus further facilitating the ligand-independent *trans*-phosphorylation of the TK domains.

Chapter 3

Methods

A perfect biophysical technique would be able to track the positions of atoms in biological macromolecules *in vivo*. It would also allow the visualization of these particles at any given time or length scale. However, such a technique does not currently exist. The current research of biomolecules is instead based on numerous different experimental, computational, and theoretical methods that complement each other by covering their respective ranges of the time and length scales [103, 104].

This Chapter provides a brief overview of the modeling and simulation methods employed in this Thesis. It also introduces the simulation models and analysis tools used to perform this research. For sufficient background, the Chapter begins by reviewing selected experimental techniques used in biophysical protein research. Namely, these experimental techniques provide the structural data used in modeling and simulations. Their ouput also serves as a crucial source of comparison for biomolecular simulations.

3.1 Selected Experimental Techniques for Studying Proteins

There are several experimental methods to explore proteins. The basis of these methods lies in genetic engineering techniques designed to express the protein of interest in sufficient quantities [1]. The following steps typically involve detection, isolation, and purification of the desired protein species. After these steps, the protein can be structurally analyzed. This Section reviews the experimental methods for characterizing the structure and function of proteins.

3.1.1 Protein Structure Determination in a Nutshell

Knowing the atomistic structure of a particular protein allows one to study its atomic-scale properties. It is also a prerequisite for the rational design of drugs that activate or inhibit the function of the protein. Ever since the publication of the DNA double helix in 1953 [105] and the first high-resolution protein structure of myoglobin in 1960 [106], the number of resolved protein structures has been growing exponentially [107, 108]. Currently, RCSB Protein Data Bank (PDB, rcsb.org) is the largest and most popular online database for 3D structural data of biomolecules, mainly proteins [109]. These experimentally resolved structures are freely available for researchers. They can be downloaded as PDB files which use a specific text-based file format for storing the atomic coordinates of the molecules. Each structure is also assigned a four-character alphanumeric identifier called PDB ID.

X-ray crystallography

Protein structure determination is dominated by two established methods: X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy [110]. Today, most of the available protein structures on PDB are determined with X-ray crystallography. This technique uses beams of X-rays to resolve the molecular structures inside a crystal [104]. The wavelength of these X-rays is ca. 1 Å, which equals the length of typical atomic bond lengths in molecules, making it ideal for scanning the relative distances between covalently-bound atoms. The basic principle is that the crystallized structure causes the beam of X-rays to diffract, producing an interference pattern that can be interpreted to construct a structural model. X-rays are scattered by electrons, so measuring the angles and intensities of the diffracted beams produces a 3D electron density map of the target molecule. Properties like mean positions of the atoms can then be deduced from the electron density map using sophisticated algorithms. The level of resolution depends on factors such as the quality of the crystals and intensity of the applied radiation.

Nuclear Magnetic Resonance Spectroscopy

NMR spectroscopy is another experimental technique for solving the 3D structures of proteins and other biomolecules [104]. It is usually done to samples in solution, thus removing the necessity for crystallization. NMR spectroscopy is therefore especially suited for resolving the structures of disordered proteins, or even some carbohydrates. In this method, the sample

is placed inside a strong magnetic field, which aligns the magnetic momenta of the NMR active atomic nuclei with the magnetic field. The sample is then exited to nuclear magnetic resonance with radio waves that match the resonant frequency of the nuclei to be measured. After the excitation pulse is over, the nuclei relax and emit the radiofrequency radiation they had absorbed in the excitation phase. This emission is detected with sensitive receivers. The frequencies being absorbed/emitted by the nuclei depend on their chemical environment, giving each nuclei a unique signal.

Structural determination of biomolecules using NMR is possible because atomic nuclei are coupled to each other via hyperfine interactions. Magnetic dipole–dipole interaction, also known as dipolar coupling, refers to the direct interaction between two magnetic dipoles. It depends only on known physical constants and the inverse cube of the distance between the nuclei, thus allowing one to determine the relative distances between different atoms. Spin–spin coupling, also known as J-coupling, refers to the indirect interaction between two nuclear spins. It arises from interactions between the nuclei and local electrons. J-coupling can provide information on the connectivity of chemical bonds as well as relative bond distances and dihedral angles. With proper experimental distance and/or dihedral restraints — and with the basic knowledge of properties, such as peptide bond geometry and bond lengths — protein structures can then be solved from the NMR data using the so-called distance geometry algorithms.

In structural studies of proteins, one typically acquires the NMR data by conducting multiple different types of NMR experiments that each reveal their own sets of spatial information about the sample in question. These experiments include two-dimensional homonuclear magnetic resonance-based methods, such as correlation spectroscopy (COSY), total correlation spectroscopy (TOCSY), and nuclear Overhauser effect spectroscopy (NOESY). Yet another common method is the nitrogen-15 NMR, including the heteronuclear single quantum coherence spectroscopy (^{1}H-^{15}N HSQC). It is currently the most widely used NMR method for resolving the solution structures of proteins.

Issues with Membrane Proteins

Integral membrane proteins possess large hydrophobic regions, which are intimately associated with the surrounding lipids. Hence, these proteins are challenging to study with classical biochemical methods in the absence of their natural lipid-rich environment. Their purification and crystallization, in particular, has proven to be extremely difficult [1]. As a result, the structural determination of membrane proteins has lagged behind that of

the soluble proteins [107]. Currently, there is a difference of two orders of magnitude in the number of resolved structures between soluble and membrane proteins, with over 100 000 available structures of soluble proteins and only less than 1000 structures of membrane proteins [111, 112]. Although full-length structures are often missing, the structures of the water-soluble domains of several membrane proteins have been resolved. In many cases, these partial structures can still help to construct an image of the full protein or its functions. There are also continuous efforts to design novel detergents that allow a better purification of membrane proteins [113,114].

Cryogenic Electron Microscopy

Recent improvements in cryogenic electron microscopy (cryo-EM) can offer help in the structural studies of membrane proteins [115]. In cryo-EM, samples are cooled fast to cryogenic temperatures and then viewed in their natural state using an electron microscope [104]. In practice, the cryo-EM samples are first embedded in vitreous water to preserve and protect them. They are then plunged into liquid ethane to rapidly freeze the water. The frozen samples are viewed by directing a high energy electron beam into them. The image forms as the electrons interact with the matter, analogous to photons in a light microscope. The rapid freezing prevents the formation of crystalline ice, which would absorb the electron beam and thus skew the final image. Overall, generating a structural representation of the sample molecule requires multiple snapshots from various angles, as well as additional image processing, to form a 3D view from a series of 2D images.

One of the advantages of cryo-EM is the lack of crystallized samples [104]. The obtained images thereby reflect the natural state of the studied molecule under the assumption that the above described procedure successfully traps the molecule in such a state. Furthermore, with this technique, one can view entities of various sizes, ranging from macromolecules to entire cells. Cryo-EM also allows controlling the chemical environment of the sample, such that it can be viewed in its physiological state. Examples of cryo-EM resolved structures include the Zika virus envelope and roughly 500 kDa bacterial β-galactosidase protein resolved with 3.8 and 2.0 Å resolution, respectively [116, 117]. While the technique has been rapidly developing over the recent years, it has disadvantages, too. Namely, it has a poor signal to noise ratio, and it is both expensive and time-consuming to use.

3.1.2 Selected Techniques for Studying the Kinetics of Membrane Proteins

Exploring the localization and diffusion of proteins on surfaces like biological membranes is typically based on fluorescence techniques, where a single particle or an ensemble of particles are labeled with fluorophores, excited, and tracked.

Total Internal Reflection Fluorescence Microscopy

When studying the membrane binding of proteins, the background fluorescence of the unbound fluorophores often masks that of the bound ones due to a large difference in their concentrations. Total internal reflection fluorescence microscopy (TIRFM) is a single particle tracking method with which a thin region of a larger sample can be visualized, thus allowing a selective excitation of the surface-bound fluorophores [118]. The method is based on the total internal reflection of incident light at a glass–water interface. It is appropriate for studying proteins in live cells with spatial resolutions of a few nanometers and temporal resolutions of dozens of milliseconds.

Fluorescence Recovery After Photobleaching

Fluorescence recovery after photobleaching (FRAP) is an ensemble-based fluorescence technique for determining the kinetics of diffusion in live cells or tissues [119]. It is especially useful in studies of protein binding and diffusion in a lipid bilayer [120]. The method is based on labeling a significant fraction of the studied molecules with fluorophores to produce a laterally uniform fluorescence intensity and measuring its recovery after a micrometer-sized fraction of the surface is photobleached.

3.2 The Need of Computer Simulations

Soft matter systems are typically tiny and fragile [7]. They are also complex, meaning they interact in multiple ways and possess various degrees of freedom, which often results in emergent and hard-to-predict behavior. Studying these systems using only experimental methods can, therefore, be challenging. For example, structural biological methods — while providing the necessary basis for current biophysical research — reveal only a static image of the molecule under study that lacks all or most of its dynamics [104]. Furthermore, crystallography methods require the crystallization

of the sample, which typically alters its physiological state. NMR spectroscopy, on the other hand, is limited by the size of the sample, as proteins above 35 kDa start to be challenging to resolve by this method [121]. Other biophysical methods, like the single-molecule tracking techniques, are good at monitoring the macroscopic motions of biomolecules, such as their localization inside a cell or in a lipid bilayer. They are, however, limited by their poor spatial resolution, so they provide very little atomistic information about the studied phenomena.

Computational methods can overcome some of these limitations. Indeed, the growing computing capacity and improved software optimization have rendered modeling and computer simulations increasingly appealing tools to study complex biomolecular systems [103]. Modern simulation methods allow the user to track the movement of biological macromolecules *in silico* with the precision of a single atom. They, therefore, complement the experiments by probing time and length scales difficult to access using other techniques. In the big picture, computer simulations can help other methods when identifying how the biological nanomachines operate and what makes them to malfunction in a disease. This, in turn, enables us to design novel treatment strategies against those pathologies.

Computer simulations and models have their limitations, too [122]. The two most important pitfalls are related to the sampling of the studied phenomena and accuracy of the used models. When can one know that the simulation is long enough, *i.e.*, the sampling is adequate? How closely does the used model mimic reality? These are questions that any computational scientist should bear in mind when designing their research. Hence, these questions will also be adressed further in the sections below.

Biomolecular simulation engines are constantly being developed and updated. As with any software, there can be bugs that cause a simulation program to produce incorrect or unexpected results. Furthermore, the used algorithms and their implementation vary between different software packages, meaning that their mutual compatibility is not guaranteed. It is, therefore, critical that computational modeling and simulation tools, codes, and results are double-checked and compared to earlier data. To this end, biomolecular simulations are typically done in close unison with related experiments to constantly verify and improve the used models.

3.3 Modeling of Biomolecules in a Nutshell

Computational research is based on models. They are mathematical descriptions generated to make better sense of the studied systems. In other

words, a model is a simplification of real-world object or phenomenon that describes only the most relevant degrees of freedom of the system under study. For example, in this work, we do not consider the behavior of entire organisms, cells, or even organelles. Instead, our focus lies on tiny model systems a few tens of nanometers in size. They describe a patch of the cellular environment, such as a membrane protein embedded in a simplified lipid bilayer or two proteins interacting in a slab of water. This is a viable research approach as long as the questions and hypotheses are appropriately placed. For example, sometimes modeling and simulating even a single domain of a larger protein is enough if it is essential for the research question, biologically justified, and verifiable *via* experiments.

Model as a word can, in principle, mean multiple different things. To avoid confusion, the nomenclature used in this Thesis is defined as follows: *model system* refers to the molecules assembled inside the simulation cell; *model* itself entails the interactions, algorithms, and equations used to capture the behavior of the model system; *simulation* means the actual running of this model with a simulation algorithm; *modeling* refers to the process of generating a model system or model, which can mean, *e.g.*, the construction of an appropriate starting structure of a protein.

Homology Modeling Helps to Construct Protein Structures

Experimentally-derived protein structures extracted from, *e.g.*, PDB are not always intact or otherwise appropriate. Depending on the resolution they were resolved with, they might lack hydrogens, side-chain-atoms, or entire regions of the protein. For example, the most flexible loops are disordered in the crystal, and hence they are often not resolved in the structure. These typical shortcomings in the available structures mean that one has to refine the coordinates of the protein to include all the necessary atoms before a simulation can be initiated. In some cases, the structure of the desired protein might not be resolved at all. In such cases, one can opt to use homology modeling to obtain the desired protein structure, provided that a sufficient, evolutionally-related template structure is available.

Homology modeling refers to the construction of an atomistic model of a target protein based on a sequence alignment between the target and a structurally-resolved, homologous template. Homology modeling relies on the fact that structures are generally more conserved than sequences among homologous proteins [123–125]. Typically, primary sequences with more than 40 % sequence identity to the target are regarded as good templates. For conducting homology modeling, there is a plethora of protein modeling software from which majority are available as downloadable programs [126]

or automated webservers [127, 128].

In this Thesis, we conduct homology modeling of proteins using a software called MODELLER [126, 129]. It is one of the most popular protein modeling tools available, and it is known to perform well in recent benchmark tests [130, 131]. One of its main advantages is that it is based on the Python programming language [132], allowing it to have bindings to various other modules and dynamic libraries. For homology modeling, MODELLER uses an automated modeling protocol that also allows for manual intervention, such that one can, *e.g.*, impose experiment-based constraints to guide the modeling [129]. It also has additional features, such as multiple alignments of protein sequences, *de novo* modeling of flexible loop regions, ability to mutate protein residues, and structure optimization based on an objective function.

In practice, MODELLER follows the general workflow of homology modeling [129]. First, it requires three things as input: target sequence, target–template sequence alignment, and template structure. It then uses a model generation algorithm called satisfaction of spatial restraints to generate a three-dimensional structure of the target protein. Based on the target–template alignments, the algorithm constructs a set of geometrical criteria and converts them to probability density functions for the location of each heavy atom in the protein. These restraints then serve as a basis for several global optimization cycles that use the conjugate gradient energy minimization algorithm to refine the coordinates of the heavy atoms [129]. Finally, the user needs to carefully assess the viability of the modeled structure by comparing it to the template. Additionally, when finally simulating the model structure, it is advisable to run a parallel control system based on the coordinates of the template structure for comparison.

Imposing Modifications to Protein Structures

Often in protein research, the key questions are related to the chemical alterations of proteins, such as mutations or PTMs, and their implications to function and disease. In case of glycosylations, complete glycan structures are rarely available beyond the first or second glycan unit attached to the protein. That is, beyond the first glycan units, the structures are difficult to resolve experimentally due to their flexible and dynamic nature [7, 27]. Hence, to make structures of glycoproteins more available for researchers, tools have been developed to *in silico* glycosylate existing protein structures. These tools include web-based portals, such as the Glycosciences web-page [133] and the CHARMM-GUI portal [134], as well as standalone programs, such as the `do_glycans` tool [135] or the Glycosylator frame-

work [136].

The paradigm of imposing chemical alterations to protein structures *in silico* differs drastically from that used in nature. *In vivo*, mutations form at the DNA level, and PTMs are done after translation, often while the protein is still undergoing folding to its native conformation. *In silico*, chemical alterations are always implemented on the native, folded proteins. It is, therefore, crucial to ascertain that the imposed changes are biologically reasonable, such that they would not, *e.g.*, affect the folding of the protein. The best way to obtain this confirmation is to have experimental data for comparison.

3.4 Molecular Dynamics Simulations

Molecular dynamics (MD) is an established computational simulation method based on classical mechanics [137]. It presents an appealing method for studying biomolecular systems because it can provide both structural and dynamic data at the atomistic level. In the MD method, molecules are described as particles attached by springs, whose dynamics are modeled with Newton's equations of motion. The result of these computations is the time evolution of the coordinates and velocities of the constituent particles. It is called the simulation trajectory. With this simple workflow, the MD method can model and predict the equilibrium properties of any classical many-body system.

Despite the straightforward idea behind the MD simulation method, one has to be aware of its limitations when evaluating its applicability to a given problem or interpreting results produced by it [137]. For instance, the classical nature of the MD method means that it neglects the electronic degrees of freedom. That is, the quantum-mechanical (QM) phenomena, such as chemical bond breaking or forming, are out of the scope of MD simulations. This lack of QM phenomena effectively sets a lower limit for the size of the system reasonable to probe with this method at the level of individual atoms — *i.e.*, the nanoscale. The maximum timescales are, in turn, limited by practical considerations, such as the size of the system (*i.e.*, number of particles), length of the time step, and available computing capacity.

Simulating biological systems is a constant balancing act between the available resources and the desired accuracy of the models [122]. Simplifying the system under study — *e.g.*, by focusing on just a small snippet of it or by changing the simulation model — will increase the available timescales, but at the cost of losing potentially important information [137]. As a

rule of thumb, the duration of a simulation needs to be significantly longer than the characteristic timescale of the studied phenomenon. These aspects have to be properly accounted for when designing MD simulation studies.

Practically, MD simulations require four things [138]. First, the simulation begins from a starting structure of the model system. Second, the user needs to select an appropriate force field to describe the properties and interactions of the modeled particles. Third, one needs to run the simulation with a simulation engine, a program that executes the production algorithm. Lastly, parameter files contain information about the physical conditions of the simulation. This information includes the choice of timescale, time step, frequency of output, temperature control, pressure control, initial velocities, and other simulation parameters.

3.4.1 Initial Conditions Define the Starting Point for a Simulation

Before starting an MD simulation, one needs to construct a system containing the studied molecules [138]. In practice, the required information comprises the coordinates of the constituent atoms and their molecular topologies. The initial coordinates of complex molecules, such as proteins, are typically based on experimentally-determined structures and acquired from online databases, such as PDB [109]. The user builds the initial structure of the entire model system by inserting the constituent particles, including solvent molecules and ions, into a simulation box. Next, one must construct the molecular topologies of the particles by defining their basic properties, such as mass and charge, as well as connections to other particles, such as bonds and angles. This information then remains fixed during the MD simulation, meaning that bonds between atoms do not break or form.

Recently, the construction of simple simulation systems has become more straightforward, due to the increase in computing power [8], availability of high-resolution structures of biomolecules [109], and the emergence of more user-friendly tools [139]. For example, model systems can be readily built using web-based portals that output simulation-ready 3D structures, topologies, and parameter files [140–145].

After defining the initial structure and parameters, one needs to verify that there are no excessive forces and that the structure is reasonably close to its equilibrium state [138]. Otherwise, reaching the equilibrium during the actual production simulation might take time. The forces in the initial structure are therefore minimized using an optimization algorithm, such as the steepest descent or conjugate gradient [138]. After these steps, the system is generally ready for the production simulation runs, where it starts

to evolve in time as the forces act on the particles.

3.4.2 Force Field Determines the Molecular Interactions in Simulations

Force field is a central concept in MD simulations, as it determines the forces that act on the particles [137]. It entails two parts: the functional form of the potential energy function and the set of parameters for different interaction terms in the function. These parameters typically include atomic masses, partial charges, van der Waals radii, as well as equilibrium values and force constants of bond lengths, angles and dihedrals between two, three, or four bonded atoms. They are typically derived for a given potential energy function using either quantum-chemical calculations or experiments, such as NMR or vibrational spectroscopy. However, there is no systematic way to find optimal values for these force field parameters.

In practice, the selection of the force field is usually one of the first tasks in the MD simulation workflow. In the field of biomolecular simulations, there are numerous force fields using their own functional forms and related sets of parameters. Therefore, the choice of an appropriate force field frequently poses an issue to a computational scientist. While every parameter set aims to model the reality as well as possible, there is no unified model that would be universally good at describing all biomolecules [146]. Furthermore, most of the existing simulation models are not mutually compatible. One must, therefore, be aware of the properties of different force fields and know how each of them is parametrized [122]. Lastly, it is vital to compare the obtained simulation results to both earlier simulations and available experimental data [103].

In the MD simulation algorithm, the forces $\mathbf{F}$ acting on a particle i are calculated by taking the negative gradient of the potential energy

$$\mathbf{F}_i = -\nabla \mathcal{H}_{\text{Total}}, \tag{3.1}$$

where $\mathcal{H}_{\text{Total}}$ is the total potential energy function [138]. As noted above, this potential energy can be divided into different interaction terms that are grouped into bonded and non-bonded functions, such that $\mathcal{H}_{\text{Total}} = \mathcal{H}_{\text{Bonded}} + \mathcal{H}_{\text{Non-bonded}}$.

Bonded Interactions

An MD force field describes covalent bonds with three bonded interaction terms: bond stretching between two particles, angle bending between three

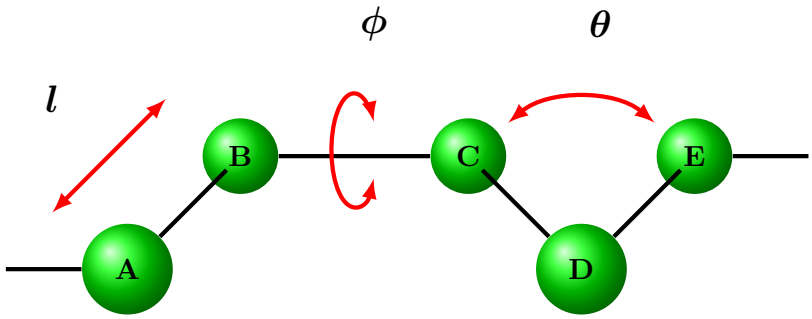

Figure 3.1: *A schematic illustration of a subsection (atoms A, B, C, D, and E) of a molecule and its internal degrees of freedom in MD. The bond stretching motion l is depicted as the relative movement between atoms A and B. Dihedral angle rotation ϕ is the angle between planes formed by atoms ABC and BCD, while θ marks the bending of the angle formed by atoms CDE.*

particles, and dihedral twisting between four particles (Figure 3.1). Additionally, the dihedral term is further divided into two subcategories. First, the proper dihedral twisting term describes rotation around a bond [138]. Second, the improper dihedral twisting term describes the maintenance of planarity of four atom groups, such as the chirality of stereospecific centers or *cis–trans* isomerism around a double bond. The potentials for these interactions are written as follows:

$$\mathcal{H}_{\text{bond}}(r_{ij}) = \frac{1}{2}k_{r,ij}(r_{ij} - r_{ij}^{\text{eq}})^2, \tag{3.2}$$

$$\mathcal{H}_{\text{angle}}(\theta_{ijk}) = \frac{1}{2}k_{\theta,ijk}(\theta_{ijk} - \theta_{ijk}^{\text{eq}})^2, \tag{3.3}$$

$$\mathcal{H}_{\text{dihedral}}(\phi_{ijkl}) = k_{\phi,ijkl}(1 + \cos(n\phi_{ijkl} - \phi_{ijkl}^{0})), \tag{3.4}$$

$$\mathcal{H}_{\text{improper}}(\xi_{ijkl}) = \frac{1}{2}k_{\xi,ijkl}(\xi_{ijkl} - \xi_{\text{eq}})^2, \tag{3.5}$$

where r_{ij} is the distance between particles i and j, θ_{ijk} is the angle between particles i, j, and k, and ϕ_{ijkl} and ξ_{ijkl} are the angles between planes ijk and jkl (zero when particles i and l are in the *cis* conformation). Parameters $k_{r,ij}$, $k_{\theta,ijk}$, $k_{\phi,ijkl}$, and $k_{\xi,ijkl}$ denote the force constants while r_{ij}^{eq}, θ_{ijk}^{eq}, ϕ_{ijkl}^{0}, and ξ_{eq} correspond to the equilibrium or reference values for bond, angle, proper dihedral, and improper dihedral, respectively. Lastly, integer n describes the periodicity. The force constants and reference values for bond-stretching and angle bending are normally drawn from experimental data. The parameters for proper dihedrals are normally obtained from fits to QM calculations.

The highest-frequency motions (*i.e.*, the bond vibrations) are described

by the harmonic potential in Equation 3.2. To improve sampling, they can be kept constant with a constraint algorithm. In practice, one typically fixes the bond lengths of all covalent bonds involving hydrogen atoms to a constant value. This treatment removes unnecessary degrees of freedom and thus allows the use of a longer time step. In this work, we kept the hydrogen-involving bonds fixed with the LINCS (Linear Constraint Solver) algorithm [147].

Non-bonded Interactions

Non-bonded interaction terms describe the interaction between two particles not connected *via* covalent bonds [138]. These terms include Pauli repulsion, van der Waals interaction between any two particles, and electrostatic interaction between two charged particles. The first two terms are described with the Lennard-Jones (LJ) potential:

$$\mathcal{H}_{\mathrm{LJ}} = 4\epsilon_{ij}\left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}}\right), \tag{3.6}$$

where ϵ_{ij} is the depth of the potential well between particles i and j, σ_{ij} is the distance at which the potential between these particles is zero (*i.e.*, the size of the particle), and r_{ij} is the distance between the particles. Values for these parameters are obtained through fits to experimental fluid properties, such as molecular volumes or heats of vaporization [148].

The basis for the repulsive $1/r^{12}$ component of the LJ potential lies on the Pauli exclusion principle, which states that identical fermions cannot occupy the same quantum state. Hence, particles repel each other at close distances. The attractive $1/r^6$ component, on the other hand, stems from the dispersion forces between two dipoles at longer distances. The decay of this attrative term with distance has its roots in physics, as freely rotating induced dipole–induced dipole interaction decays with $1/r^6$. The exponent of the repulsive term $1/r^{12}$ has, however, been chosen only for simplicity and computational efficiency, while retaining a sufficient physical accuracy.

Electrostatic interaction is described with the Coulomb potential

$$\mathcal{H}_{\mathrm{Coulomb}} = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}}, \tag{3.7}$$

where q_i and q_j are the charges of particles i and j, r_{ij} is the distance between these two particles, ε_0 is the permittivity of vacuum, and ε_r is the relative permittivity of the medium. In MD simulations, each particle is usually assigned a partial charge, which is normally derived from QM

calculations, to realistically describe the distribution of electrons in the entire molecule.

In practice, it is costly to calculate long-range non-bonded interactions to every molecule of the system. Hence, the MD algorithm employs cut-offs, *i.e.*, it omits the non-bonded interactions after a certain threshold distance [137]. The LJ interaction is typically cut at 1 nm, yet the loss of potential energy is negligible due to its rapid decay with distance. The Coulombic interaction, however, remains significant at longer distances, causing major computational load for the simulation. Therefore, the Coulomb's law is usually used directly until ca. 1 nm, after which the long-range electrostatic forces are calculated by interpolating the charge density into a discrete grid in the reciprocal space and evaluating it using the Fast Fourier transform. This protocol is called the Particle Mesh Ewald (PME) method, and it is commonly used in the field of MD simulations due to the enhancement in the runtime it offers [149]. Finally, to keep track of the neighboring particles within the cut-off radius, MD simulations use neighbor lists [137]. This relatively costly operation is accelerated by updating the lists only at defined intervals — *i.e.*, not at every simulation time step.

Biomolecules Can Be Modeled with Varying Levels of Detail

In biomolecular research, there are two basic types of classical force fields: atomistic and coarse-grained [103,137]. In atomistic force fields, each simulated particle describes one atom, whose parameters are derived from experiments, as explained above. These force fields, therefore, reflect reality as well as possible. They can, for example, accurately describe biologically relevant atomic-scale events, such as conformational dynamics of amino acid side-chains or hydrogen bonding dynamics. Atomistic force fields can be further divided into two: all-atom (AA) force fields that provide parameters for every atom in the system and united-atom (UA) potentials that spread the contribution of non-polar hydrogens into their host heavy-atoms [150]. UA models accelerate the simulation compared to AA models but loose some of the atomistic precision, as there is a fewer number of particles to simulate.

The AA description is often the natural choice for modeling biomolecular systems if the focus lies on the structural details. Popular AA force fields used in the MD simulations of biomolecules include AMBER ff99SB-ILDN [151], CHARMM36 [152,153], GROMOS [154], and OPLS-AA [155]. The research in this Thesis employs the first two of them (AMBER ff99SB-ILDN, CHARMM36), mainly due to their documented performance and their extensive selection of parameters for various molecule types, includ-

ing proteins and glycans [134, 146, 151]. The AMBER ff99SB-ILDN description is especially suited for proteins but is also compatible with the GLYCAM06 [156] parameter set for carbohydrates and Lipid14 [157] parameters for lipids. Likewise, the AA CHARMM36 protein force field [152] can be combined with the recently-developed CHARMM36 lipid and carbohydrate force fields [158–160].

As noted above, atomistic models form the basis of MD simulation-based research of biomolecules. However, the sampling of the relevant length and timescales with atomistic models is sometimes inadequate to describe large-scale phenomena, such as the binding of a protein complex to a lipid bilayer or dynamics of the glycocalyx [103]. Coarse-grained (CG) force fields improve this sampling while still retaining some general level of the chemical specificity [138]. In practice, CG models simplify the system by mapping multiple AA heavy atoms into pseudo-atoms or beads. Details of the mapped particles are then smeared into the parameters of the bead. This treatment provides ca. a ten-fold reduction in the number of particles in a given system. The consequent decrease in the number of degrees of freedom accelerates the simulation significantly [161, 162]. Additionally, losing the fastest degrees of freedom, such as bond vibrations, allows the use of longer simulation time step, which further expedites the simulation.

The most common CG force field in biomolecular simulations is the MARTINI model [161, 162]. Its popularity stems from its extensive selection of molecule types [161, 163, 164] and well-documented tools [139] that render it effortless to use. Furthermore, it is implemented on three major MD simulation software: GROMACS [8, 165], GROMOS [166], and NAMD [167]. MARTINI is predominantly based on four-to-one mapping, in which one bead represents four heavy-atoms [162]. Currently, the model has 20 bead types, divided into four main bead types with each having several subtypes, to describe the chemistry of typical biomolecules including proteins, carbohydrates, lipids, and nucleic acids. It also supports both implicit and explicit solvent models. MARTINI is parametrized based on thermodynamic properties, mainly partitioning free energies between oil and water [161]. It thereby combines reasonable structural accuracy with thermodynamic reproducibility.

Like any modeling, coarse-graining has its pitfalls. When disregarding microscopic degrees of freedom, one has to make sure that they are not relevant to the studied research question. For example, atomistic details, such as hydrogen bonds cannot be reproduced by highly coarse-grained models. Additionally, the physical basis for any observable may be altered due to coarse-graining, leading to a wrong interpretation of the underlying

mechanism. Furthermore, coarse-graining causes entropy loss as molecules become more simple, which may cause the relation between entropy and energy to be unphysically balanced [168]. MARTINI is also known to suffer from model-specific issues, such as the overbinding of proteins [169]. One way to overcome these limitations is to combine AA and CG models in multiscale modeling, where the same system is simulated with both models in an alternating fashion. Indeed, recent tools have allowed for a straightforward interconversion between the two descriptions. For example, the `martinize` tool converts AA structures of proteins into the MARTINI CG representation, while the `backward` tool reverses this by converting CG models back to AA representation [170].

3.4.3 The Molecular Dynamics Simulation Algorithm

Classical MD simulations model the dynamics of particles by solving Newton's second equation of motion [137]. The algorithm begins by calculating the forces acting on the constituent particles by taking the negative gradient of their potential energy function. When it knows the forces acting on the particles, it will update their positions and velocities from their previous values through numerical integration. The algorithm then repeats these steps over multiple discrete time steps to move the atoms. For the stability of the simulation, the user typically chooses the time step to be short — *i.e.*, not more than two femtoseconds. In practice, simulations can span billions of such time steps, producing molecular trajectories that cover microseconds of the large-scale dynamics of the simulated particles. For example, small water-soluble protein domains have been simulated for 100–1000 microseconds — enough to observe protein folding [171].

In this work, we used the so-called leapfrog integrator algorithm for the numerical integration [172, 173]. It calculates the velocities at every half step

$$\mathbf{v}_i\left(t + \frac{\Delta t}{2}\right) = \mathbf{v}_i\left(t - \frac{\Delta t}{2}\right) + \frac{\mathbf{F}_i(t)}{m_i}\Delta t \tag{3.8}$$

and positions at whole steps

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i\left(t + \frac{\Delta t}{2}\right)\Delta t, \tag{3.9}$$

such that $\mathbf{v}_i$ and $\mathbf{r}_i$ denote the velocity and position of particle i, respectively. Quantity $\mathbf{F}_i(t)$ is the force acting on particle i of mass m. Quantities t and Δt are the simulation time and time step, respectively.

In practice, the numerical integration and the rest of the MD algorithm are executed by a simulation engine. Currently, there is a range of different simulation engines, and each comes with their own sets of functionalities that can differ vastly from what the other platforms have to offer [165–167]. In this Thesis, we have used the GROMACS simulation software package [8, 165]. It is among the most popular MD simulation software available. For example, it is the main MD simulation engine used by the popular folding@home project for distributed computing [174]. Its popularity stems from the fact that it is free and constantly-updated, open-source software, which can be run on both central processing units (CPUs) and graphics processing units (GPUs) [175]. It is also one of the fastest simulation software packages available with the ability to be readily parallelized using the Message Passing Interface (MPI) or threads [8]. Furthermore, GROMACS supports several AA and CG force fields and has a large variety of analysis and system preparation tools, which can be further expanded by the user thanks to its open-source nature. The software lacks an in-built molecular viewer, but one can readily circumvent this issue by using, *e.g.*, the Visual Molecular Dynamics (VMD) software for visualizing molecular structures and trajectories [31].

3.4.4 Simulation Conditions

By default, running an MD simulation leads to the microcanonical (NVE) thermodynamic ensemble [137]. In the NVE ensemble, the number of particles, the volume of the system, and the total energy of the system remain constant over time. However, in the related real-life experiments, temperature is usually constant instead of the total energy of the system. Such a situation corresponds to either the canonical (NVT) or isothermal–isobaric (NpT) ensemble. To reach the NpT ensemble in MD simulations, one must use thermostats and barostats to keep the temperature and pressure of the system fixed. For the NVT ensemble, one needs only thermostat, as pressure control is not required.

The aim of simulation thermostats is to maintain the temperature at a fixed level, but at the same time allow it to fluctuate according to the canonical distribution [137]. A simple way to achieve this is to mimic a weak coupling of the system to an external heat path. In the so-called Berendsen thermostat, this is accomplished by scaling the velocity of each particle with a rate determined by a time constant at each time step [176]. However, this suppresses the fluctuations of the kinetic energy and therefore fails to produce the canonical ensemble. As the Berendsen thermostat is nevertheless efficient in stabilizing the temperature of the system, one typically uses it

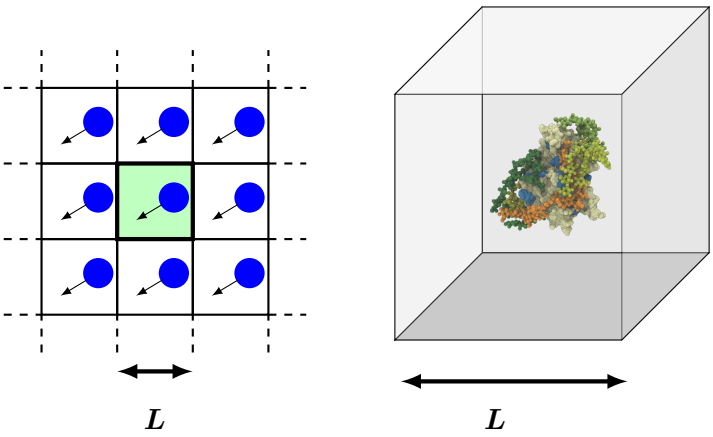

Figure 3.2: *On the left: a schematic figure representing the fundamental idea of periodic boundary conditions in 2D. The unit cell along with the particle trajectories are copied in all directions and if a particle tries to leave the cell it simply emerges back from the opposite side. The parameter L is the size of an individual cell. On the right: a real unit cell is given as a snapshot from real simulation, where the CD44 protein is placed at the center of the simulation box. Water molecules are omitted from this figure for clarity.*

to equilibrate the solvent and ions around the studied macromolecule at the early stages of the simulation workflow. The subsequent production simulations employ more sophisticated algorithms, such as the Nosé–Hoover thermostat [177], able to generate trajectories consistent with the canonical ensemble.

Typical barostat algorithms are largely analogous to the simulation thermostats, coupling the system to an external reference pressure. For example, the Berendsen barostat scales the simulation box vectors and particle coordinates that lead to first-order kinetic relaxation of the pressure towards the reference pressure [176]. However, the Berendsen barostat is unable to yield the correct isothermal–isobaric ensemble and is thus used exclusively in the equilibration phase. The production runs typically employ the Parrinello–Rahman barostat due to its ability to produce the desired statistical ensemble [178].

Simulation systems are small and finite. Thus, to approximate a vast, infinite system using a small unit cell, simulations use periodic boundary conditions (PBCs). They replicate the unit cell or simulation box in every direction, effectively generating an infinite system formed from copies of the original cell, as shown in Figure 3.2. During a simulation, only the coordinates in the original simulation box need to be recorded by the MD algorithm. Due to PBCs, a particle leaving the unit cell from one side re-

appears from the opposite side. A particle also interacts with the closest periodic images of the other particles in the system even if they were in the adjacent cell. Therefore, the size of the simulation box must be carefully selected to ensure that molecules do not interact with their own periodic images, as that would be unphysical. Besides enabling the simulation of bulk properties with a small unit cell, PBCs introduce translational symmetry into the simulation. This property ensures the conservation of total momentum of the system [179]. Namely, in a finite system, a wall potential would give rise to finite size effects and abrogate the symmetry as well as the conservation of momentum.

3.5 Enhanced Sampling and Analysis Methods

Once the simulation trajectory is sufficiently long, it can be analyzed in multiple ways. In this Thesis, we employ a range of analysis tools, including both standard GROMACS analysis codes and in-house scripts. The use of these tools is explained where appropriate. This Section reviews only the primary analysis techniques used in this work. The errors reported in this Thesis are standard errors unless stated otherwise.

Computing Free Energies with the Umbrella Sampling Method

The purpose of an MD simulation is to ultimately sample all the states in which the particles can exist [180, 181]. Given enough sampling, one can calculate the probability for a particle to be in each state. Such probability is, in turn, related to the free energy of that state. In practice, however, the states are separated from others by high free energy barriers. In some cases, these barriers can be so high that the sampling of the corresponding states would be highly insufficient with regular equilibrium MD simulations [180].

Free energy methods, such as the umbrella sampling [182], allow the sampling of states with low probability by overcoming the free energy barriers that separate them. The umbrella sampling technique overcomes these barriers by imposing additional harmonic biasing potentials, thus forcing the system to explore also regions of the state space where ergodicity would otherwise be hindered. The unbiased free energy profile can then be computed by substracting the used biasing potential from the biased, umbrella-sampled free energy profile [180].

The typical biological model systems involve numerous degrees of freedom, ranging from intramolecular motions to intermolecular interactions. Hence, one typically applies the umbrella biasing potentials to only one or a few of these degrees of freedom that somehow represent the phenomena

under study. Such degrees of freedom are called reaction coordinates. The umbrella sampling technique divides the selected reaction coordinate into intermediate steps and then applies the biasing potential to each of these steps [180]. Practically, each step is a separate simulation window set to sample a particular region of the reaction coordinate.

The sampling is adequate when the system has visited every value along the reaction coordinate multiple times — *i.e.*, the simulation windows are long enough to capture the studied phenomenon more than once [180]. Combining the intermediate windows yields a biased probability distribution along the reaction coordinate, which can be unbiased to give the potential of mean force that approximates the free energy along the chosen coordinate. This unbiasing is typically conducted with the weighted histogram analysis method (WHAM) [183], which also offers a way to minimize the statistical error through an iterative process. In GROMACS, WHAM is implemented under the `gmx wham` code [184].

Estimating Binding Affinities with the Molecular Mechanics Poisson–Boltzmann Surface Area Method

When evaluating the binding affinity of two solvated molecules, one would ideally simulate their spontaneous binding to measure the corresponding free energy of the process [185]. In such simulation, however, the majority of the energy contribution would stem from solvent–solvent interactions and the fluctuations in the total energy would be orders of magnitude higher than the binding energy. These fluctuations would then lead to slow convergence of the calculation. Although traditional free energy methods, such as umbrella sampling, would in most cases be ideal for such a calculation, they often require a considerable amount of resources to accurately capture these processes. On the other hand, typical drug design-related methods, such as molecular docking and scoring [186], would estimate the binding free energy efficiently but not particularly accurately.

The so-called endpoint methods fall between the two extremes. In these techniques, only the endpoints — *i.e.*, the complex and its unbound components — are sampled, making them less expensive than umbrella sampling, but more accurate than simple docking and scoring [187]. Molecular mechanics Poisson–Boltzmann Surface Area (MM/PBSA) is a popular endpoint approach for estimating the binding free energy of two solvated molecules [185, 188]. In MM/PBSA, the free energy of binding ΔG_{bind} is

estimated by decomposing it into components according to

$$\Delta G_{\text{bind}} = \Delta G_{\text{bind,vacuum}} + \Delta G_{\text{solvation,complex}} - \left(\Delta G_{\text{solvation,subunit1}} + \Delta G_{\text{solvation,subunit2}} \right), \tag{3.10}$$

where $\Delta G_{\text{bind,vacuum}}$ is the free energy of binding in vacuum and the solvation terms $\Delta G_{\text{solvation}}$ describe the solvation free energy of the binding complex or dimer as well as the individual subunits.

The $\Delta G_{\text{bind,vacuum}}$ term can be further divided into

$$\Delta G_{\text{bind,vacuum}} = \Delta E_{\text{MM}} - T\Delta S_{\text{normal mode}}, \tag{3.11}$$

where ΔE_{MM} entails the standard MD energy terms from bonded, electrostatic and van der Waals interactions. In practice, one obtains them by performing equilibrium MD simulations [187, 188]. The last term comprises temperature T multiplied by entropy S, which is estimated by a normal mode analysis of vibrational frequencies. Normal mode analysis can, however, be computationally expensive and have a large margin of error. Hence, one can neglect the entropy term if the objective is to compare two states with a similar entropy, such as a wild-type and a single-amino acid mutant of the same protein complex.

The free energies of solvation are comprised of two components according to

$$\Delta G_{\text{solvation}} = \Delta G_{\text{polar}} + \Delta G_{\text{hydrophobic}}. \tag{3.12}$$

The ΔG_{polar} term is obtained by solving the Poisson–Boltzmann (PB) equation for each of the states (*i.e.*, the complex and both subunits). The hydrophobic term is estimated from a linear relation to the solvent accessible surface area (SASA) [188].

In this Thesis, the MM/PBSA calculations are conducted with the `g_mmpbsa` tool for GROMACS [189, 190]. It assumes a so-called single trajectory approach, in which the conformational changes that occur upon binding are considered negligible. With this assumption, uncorrelated snapshots of all the molecular states (*i.e.*, the complex and its subunits) can be obtained from a single MD simulation trajectory of the binding complex.

Analyzing Molecular Distances

The probabilities of two molecular entities to be in contact are calculated with `gmx minist` tool, which calculates minimum distances between two particles or groups of particles. After the minimum distance between the two entities is recorded, values below a certain threshold (typically 0.6 nm) are considered as being bound or interacting.

3.6 Overview of the Model Systems Studied in This Thesis

This Section describes the simulation systems explored in this Thesis. It lists the main components of the systems, describes their preparation, and gives rationale to why such model systems were considered. This Section also lists the most important physical variables used in each publication, yet one should refer to the original research papers for the complete sets of parameters.

Publication I

In *Publication I*, we provide a detailed view of the binding of HA to its CD44 receptor through unbiased equilibrium MD simulations [191]. To first acquire an idea of a spontaneous receptor–ligand binding, our model systems consisted of one CD44 HABD and one HA_{16} oligomer in an uncomplexed state, as shown in Figure 3.3A. Running a total of nine MD simulations of such systems then provided us with multiple structures of spontaneously-formed HABD–HA complexes. Before assembling the systems, we obtained the coordinates of CD44 HABD from PDB with identifier 1UUH [10]. The HA coordinates we constructed with the `do_glycans` tool [135]. When naming the HA oligomers, the suffix indicates the number of saccharide residues involved in the strand.

Based on the output of the spontaneous binding simulations, we characterized three HABD–HA binding modes, from which one is the previously-reported [11] crystallographic binding pose (Figure 3.3B). Each binding mode was then simulated further with three replica simulations. We did this to acquire further statistics of each pose and to ascertain their stability. The complexed HABD–HA structures served also as starting points for subsequent umbrella sampling simulations, which assessed the affinities of the binding modes. To set up the umbrella sampling runs, we first pulled the HA ligand from the surface of the protein to the bulk water. That is, we selected the Euclidean distance between HA and HABD as the reaction coordinate for the umbrella sampling free energy calculation. The subsequent analysis required a total of 30 umbrella sampling windows of 100 ns to construct the biased probability distribution along this distance-based reaction coordinate. Finally, we used WHAM to unbias the distribution and approximate the free energy of the binding process [184].

To simulate the multivalent binding of CD44 HABDs to a single HA polymer, we constructed systems of HA_{64} and two HABDs (Figure 3.3C). Here, we fixed the HA polymer to an elongated conformation by restrain-

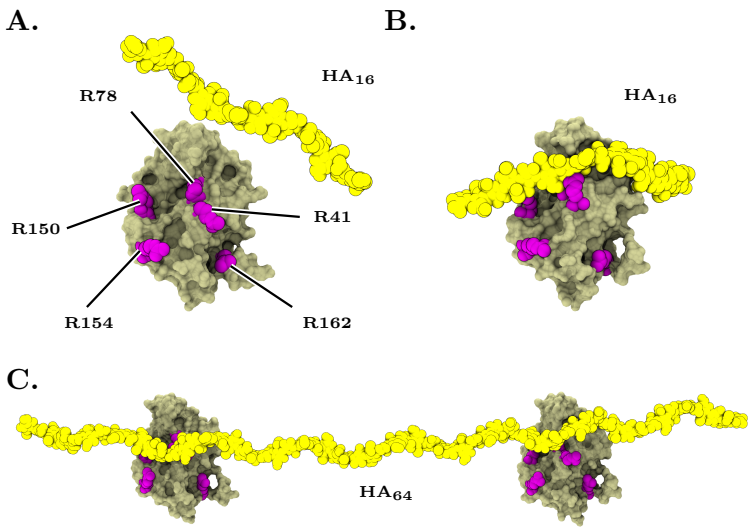

Figure 3.3: *Simulation systems in Publication I. A) Starting frame from the spontaneous binding simulations. B) Example of a CD44–HA binding complex (crystallographic). C) Starting frame from the simulations exploring multivalent binding. Note that in panels A and C HA lies in front of the proteins — i.e., it is not bound. HA is depicted as yellow, CD44 is pale surface, and the key HA binding residues are marked as purple beads.*

ing both of its tail-ends with mild (force constant of 10 kJ mol^{-1} nm^{-2}) position restraints in the direction parallel to the polymer. As a result, the HA molecule was able to wobble only along the direction of its long axis. Namely, its movement in other directions was restricted with stronger restraint potentials of 200 kJ mol^{-1} nm^{-2} to mimic a rod-like polymer. The two HABDs were also restrained from their C-terminus to mimic their attachment to the rest of the protein and to the underlying lipid membrane.

Publication II

In *Publication II*, we *in silico* glycosylated the CD44 HABD with several *N*-glycan profiles (see below) [192]. We aimed to observe how the oligosaccharides behave and fold, especially around the HA binding residues. The following Sections introduce the selected glycan profiles as well as the construction of the corresponding simulation systems.

N-glycosylations on CD44 HABD

Figure 3.4 lists all the *N*-glycan profiles, or glycoforms of CD44 HABD studied in Publication II. To model the most realistic *N*-glycosylations, we used complex type triantennary *N*-glycans, having zero or one terminal sialic acids per antenna (*i.e.*, each non-reducing glycan termini). We chose this particular triantennary structure because CD44 has been shown to bear highly-branched, complex type *N*-glycans on its surface in both wild-type and ligand-binding subclones of Chinese Hamster Ovary (CHO) cells [75]. Similar glycovariants have also been described in previous reports of the field [17]. Furthermore, a recent mass-spectrometry study demonstrates that human CD44 HABD houses triantennary complex type glycans, with a varying number of sialic acids, as the primary *N*-glycan structure at the N25-N110 positions when expressed in mouse myeloma cells [65]. The N120 site, on the other hand, houses mainly triantennary high-mannose type structures, predominantly without fucosylation [65]. Finally, to model less cancerous phenotypes, we used shorter oligosaccharides described below.

In the *full GlcNAc* glycoform, the *N*-glycans comprise only a single GlcNAc residue. We chose this glycosylation profile to evaluate the putative positive effect of the core GlcNAc residues on HA binding [17]. The *full pentasaccharide* glycoform also mimics small, neutral oligosaccharides. We chose this glycan structure as it forms the basis for all the higher-level glycans, and thus, serves as a good baseline.

Previous studies have shown that mutations N25S and N120S can convert inducible cell lines to constitutively active HA-binders [16]. To emulate such glycosylation-deficient mutant proteins, we built a *partial monosialo* glycoform of HABD, which houses monosialic complex type glycans at positions N57, N100, and N110 but lacks *N*-glycans entirely at N25 and N120.

The *full asialo* glycoform entails complex type glycans terminating to galactose residues. These glycans mimic the situation after a neuraminidase treatment of the inducible cell lines [14]. Namely, the neuraminidase enzyme cleaves sialic acid residues from the non-reducing termini of the branched glycans and has been observed to increase the binding affinity of CD44–HA complexes in several experimental studies [14, 15, 75].

The *full extended asialo* glycoform entails oligosaccharides that have two galactoses in each non-reducing glycan termini. This structure ensures the glycans are of the same size as the ones in the *full monosialo* glycoform, thus enabling the comparison of same-sized glycans (*i.e.*, –Gal–Gal or –Gal–Neu5Ac). That is, the glycan antennae of the *full monosialo* *N*-glycans terminate to a single sialic acid residue, which is preceded by a galactose residue. We chose the *full monosialo* glycoform as it corresponds to the

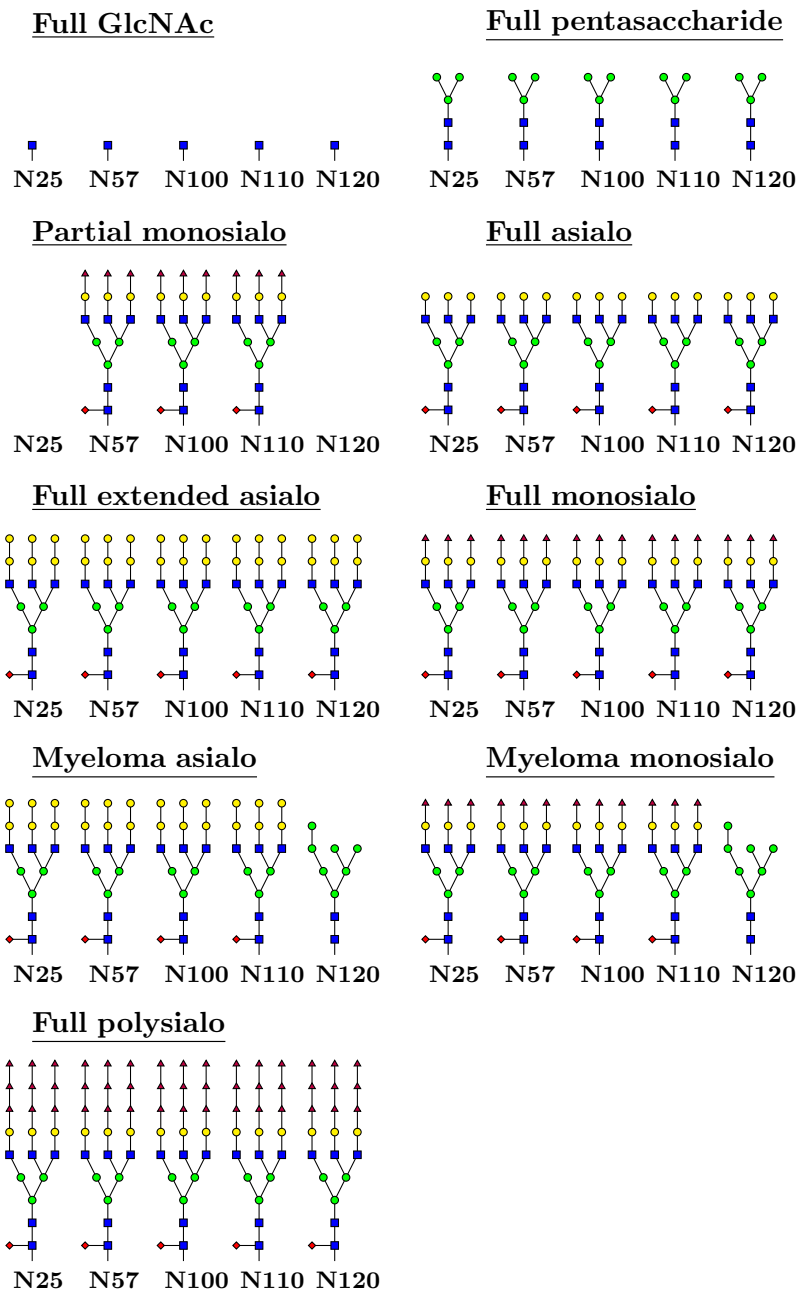

Figure 3.4: *CD44 glycoforms used in Publication II. The symbols follow the Symbol Nomenclature for Graphical Representations of Glycans [193].*

primary glycoforms found in the inducible cell types [14, 17].

In the *myeloma asialo* and *myeloma monosialo* glycoforms, glycosylation sites N25–N110 house fucosylated triantennary complex type glycans [65]. They have either two galactoses or a galactose–sialic acid pair at the end of each nonreducing terminus. The N120 position houses a non-fucosylated high-mannose structure, as described in Figure 3.4. The system name reflects the number of terminal sialic acids (*i.e.*, zero or one) per antenna on the N-glycans in glycosylation sites N25–N110.

The *full polysialo* glycoform entails polysialic complex type N-glycan structures. In this glycoform, each Neu5Ac residue connects to the next glycan residue with an α-2,3 bond. We chose these glycans to mimic the more heavily sialylated glycoforms presumed to be present in the completely inactive cell-lines. These cells do not bind HA even after a standard neuraminidase treatment [17].

System construction in Publication II

We deviced two main system setups to explore the behavior of the N-glycans. In the first setup, we modeled different glycoforms of an isolated CD44 HABD. Here, we generated three different starting configurations for the attached N-glycans. Each configuration then served as a starting structure for five replica simulations of 1000 ns each so that the total number of replicas was 15 per glycoform. The tested glycoforms included *myeloma monosialo*, *myeloma asialo*, *partial monosialo*, and *full pentasaccharide*. The HA ligand was omitted from these systems.

In the second setup, we constructed spontaneous binding systems where a HA_{18} ligand was initially placed at least 2 nm from a glycosylated HABD. This approach guaranteed a fully spontaneous binding process between the molecules without any *a priori* knowledge of factors such as binding modes. In total, we studied seven glycoforms, each having eight 1000 ns-long simulation replicas. These glycoforms included *full GlcNAc*, *full pentasaccharide*, *full asialo*, *full extended asialo*, *partial monosialo*, *full monosialo*, and *full polysialo* (see Figure 3.4). Finally, we compared the results to the analogous spontaneous binding simulations of the non-glycosylated HABD described in *Publication I*.

Publication III

In *Publication III*, we modeled an activated dimer of JAK2–TpoR complex [194]. We also simulated individual domains of JAK2 to probe their dimerization and membrane binding. To this end, we employed both AA

and CG models. The following Sections explain the construction of the simulation systems of both descriptions.

All-Atom Systems

We generated an all-atom model of JAK2–TpoR ΔECD dimer embedded into a lipid membrane. Here, TpoR ΔECD refers to a TpoR molecule lacking the extracellular domain, which was found dispensable for the formation of the active dimer complex in our related experiments (see original publication) [194]. In modeling the dimer complex, we implemented a bottom-up scheme for the system construction, as indicated in Figure 3.5. Notably, we used the known dimerization interfaces, JAK2 PK–PK (obtained with homology modeling from the X-ray structure of JAK1 [82]) and TpoR TM–TM (obtained from our simulations as well as online structure prediction tools [195]), to guide the construction of the full-length model.

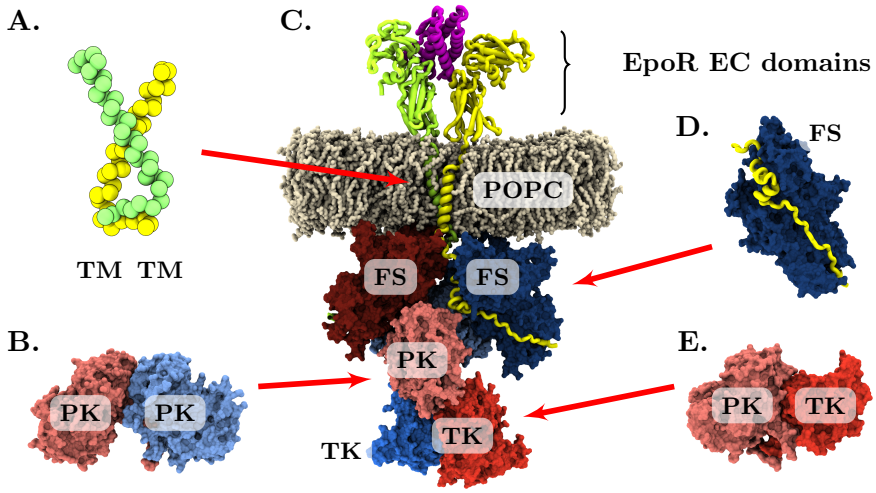

Figure 3.5: *Construction of the full JAK2–EpoR dimer model. A) CG model of the EpoR transmembrane (TM) domain dimer. B) JAK2 pseudokinase (PK) domain dimer. The two identical domains are separated by colors. C) The full receptor–JAK2 model embedded in a pure POPC bilayer (pale liquorice). The first JAK2 is depicted in shades of red and the second in shades of blue. The first EpoR is marked as light green and the second is marked as yellow. The erythropoietin ligand is colored purple. D) JAK2 FERM domain (blue) bound to the intracellular (IC) domain of EpoR (yellow). E) JAK2 pseudokinase–tyrosine kinase complex. The steps involved in the construction of these protein models are described in the text as well as in the original paper [194].*

At first, we homology modeled the PK–PK interface of JAK2 based on

the structure of the PK–PK dimer of homologous JAK1 (PDB:4L00) [82] and a native JAK2 PK domain (PDB:5UT3) [83]. We then linked the JAK2 PK–PK dimer to two JAK2 FERM-SH2 domains (PDB:4Z32) [90] that were in complex with the IC segments of TpoR. This FERM-SH2–TpoR complex was obtained by homology modeling based on the structures of JAK1–IFNLR1 (PDB:5IXD) [89] and TYK2–IFNAR1 (PDB:4PO6) [88] FERM-SH2–receptor complexes. When joining the JAK2 FERM-SH2 and PK–PK dimers, we oriented the FERM-SH2 domain such that its F2 subdomain was facing the membrane lipids, since it entails a positively charged patch of residues presumed to mediate the membrane interaction [90]. Furthermore, we placed the two FERM-SH2 domains apart, as we could not detect interactions between them. We completed the JAK2 model by modeling the TK domains based on the TYK2 TK domains within the structure of TYK2 PK–TK (PDB:4OLI) [84] and a native JAK2 TK domain (PDB:4IVA) [85]. When joining the TK domains to the PK domains, we opted to use an unstructured linker segment to mimic the extensive interdomain flexibility of JAK2 observed in EM imaging [196]. In this manner, the TK domains have sufficient flexibility in the presumed active state conformation to cross-phosphorylate each other. Finally, we linked the dimer complex to a TpoR TM–TM dimer that we obtained through sampling from our simulations.

We also constructed an analogous all-atom dimer model of JAK2–EpoR (Figure 3.5). Since the extracellular domains (EC) of EpoR have been resolved as a ligand-induced dimer (PDB:1CN4) [197], we incorporated them into the model to evaluate their possible role in the active dimer complex. In this full-length JAK2–EpoR model, the FERM-SH2–EpoR complexes were extracted from a recently-resolved tetrameric JAK2 FERM-SH2 structure (PDB:6E2Q) [81]. The EpoR TM domains were extracted from an existing NMR-resolved structure (PDB:2MV6) [198] and dimerized based on the output of the PREDDIMER web server [195]. The EpoR EC domains were taken from PDB with identifier 1CN4 [197] and linked directly to the N-terminal ends of the TM domains. Finally, the JAK2 PK and TK domains were modeled and attached as described above for the JAK2–TpoR ΔECD model. The overall molecular architecture of the JAK2–EpoR model resembles that of the earlier-built JAK2–TpoR ΔECD model. Its main differences are a slightly more tilted orientation and a surprising, EpoR-mediated dimerization of the JAK2 FERM-SH2 domains that both stem from the 6E2Q structure. Finally, we embedded both JAK2–EpoR and TpoR systems into a lipid bilayer using the CHARMM-GUI membrane builder online tool [140, 141] and simulated them for 1000 ns with multiple

simulation replicas [194].

To study the PK–PK interface more closely, we generated simulation systems of isolated PK domain dimers as described above. The aim was to focus on the molecular mechanism of the cancer-inducing V617F mutant [199–201]. Hence, we modeled the PK systems as both wild-type and V617F. From the subsequent simulations, we analyzed the differences in the binding free energies between the two mutant states using the MM/PBSA scheme [190].

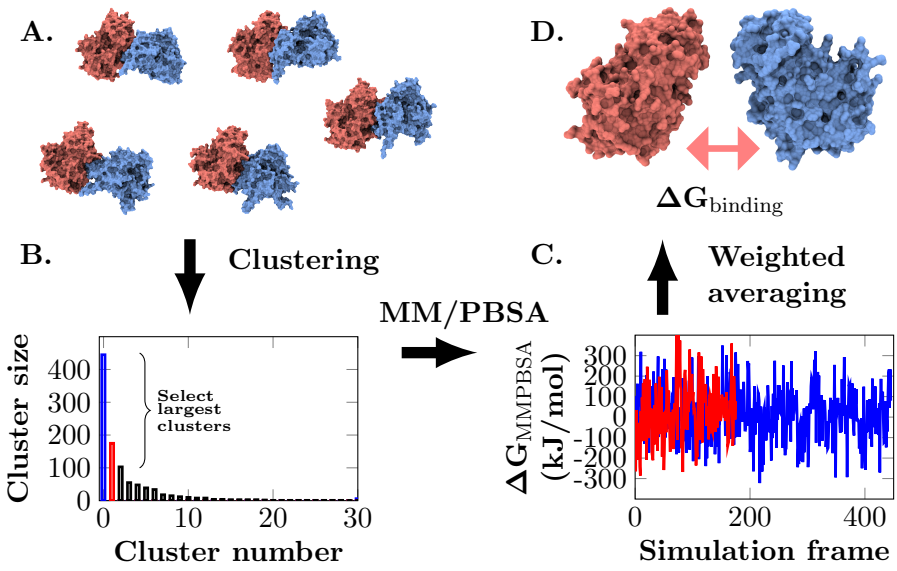

Figure 3.6: *Workflow in the MM/PBSA calculations of JAK2 PK–PK dimerization affinity. At first, all the simulation frames (A) are filtered with an RMSD-based clustering. The most populated clusters (B) are then subjected to the MM/PBSA calculation, which yields free energy values for each selected simulation frame (C). These values are then averaged to obtain an estimate of the free energy of binding (D).*

We extracted the data for the MM/PBSA analysis as uncorrelated snapshots at 1 ns intervals from the simulations and filtered them with an RMSD-based clustering using a cutoff of 0.25 nm (Figure 3.6). Two largest clusters per simulation were then selected for the final MM/PBSA analysis. This filtering procedure ensured that we were analyzing the relevant dimer poses in each case instead of the less populated ones, which occurred due to the flexible nature of the PK–PK interface [82]. In total, we simulated both wild-type and V617F states with ten 1000 ns replicas.

Coarse-grained Systems

To speed up the sampling in the simulations, we generated two system setups using a CG description. First, to determine the structure of the TpoR TM dimer, we simulated two initially unbound TM helices in POPC bilayer using the MARTINI model [162, 163]. Here, we considered both wild-type and the oncogenic W515L mutant. The simulation results showed an X-shaped dimer to be the most common structure, and as this structure also agreed with the output of the PREDDIMER tool [195], it was selected as the starting TM–TM dimer configuration for the JAK2–TpoR ΔECD system.

Second, we simulated JAK2 FERM-SH2–TpoR ΔECD monomers in a POPC/POPS bilayer (intracellular leaflet contained 10 % of POPS) using the MARTINI description [162, 163]. We explored this system in three mutant states: wild type, L224A, and L224E. The goal was to evaluate the role of the L224 residue in membrane attachment over long timescales (20 μs).

Models & Simulation Parameters

In *Publications I–II*, we modeled the dynamics of proteins with the AMBER ff99SB-ILDN force field [151] and carbohydrates with the GLYCAM06 description [156]. To acquire a more comprehensive view of the binding, we also replicated selected systems using the CHARMM36 force field [153]. In *Publication III*, we used the CHARMM36 force field for the AA systems and MARTINI 2 for the CG systems [162, 163]. Water molecules were described with the TIP3P water model [202] in all the AA systems of this Thesis. Sodium chloride (NaCl) was also added to neutralize the systems and to reach the physiological saline concentration (150 mM).

All the proteins considered in this Thesis are human homologs. When constructing these protein systems, possible missing atoms were filled in with the MODELLER tool [126]. In these cases, the canonical sequence templates were taken from the UniProt database [203, 204]. Following the system construction, every CHARMM36/MARTINI simulation was initiated with the CHARMM-GUI tool [143]. In other cases (AMBER ff99SB-ILDN/GLYCAM06), the system construction was done manually or using in-house tools [135]. *Publications I–II* were run with the GROMACS version 4.6.7 [205] and *Publication III* with the version 2016.

Every simulation was conducted in the NpT ensemble at 310 K temperature and 1 bar pressure. The temperature was coupled separately for proteins, solvent, *N*-glycans (*Publication II*), HA (*Publication I–II*), and

membrane (*Publication III*) using either v-rescale [206] (*Publication I–II*) or Nosé–Hoover [177] (*Publication III*) thermostat with a time constant of 1 ps. The pressure was coupled isotropically in membrane-exclusive systems and semi-isotropically in the membrane-inclusive systems using the Parrinello-Rahman barostat [178]. The simulations were run with the leap-frog integrator with a time step of 2 fs. PBCs were applied in all three spatial dimensions. All the covalent bonds involving hydrogen atoms were constrained with the LINCS algorithm [147]. Trajectories were saved at every 100 ps. For further details, please refer to the original publications.

Chapter 4

Binding of CD44 Receptor to Its Hyaluronan ligand

The structure of a HABD–HA binding complex has been resolved with X-ray crystallography [11]. This structure defines the binding of HA to the binding groove on the Link module. However, other studies employing mutagenesis and NMR spectroscopy have identified HA binding residues and residue clusters outside this binding groove region [21–24]. Most notably, these studies have mapped prominent HA binding epitopes to arginine and lysine-rich loci both in the Link module (R29, K38, R41) and in the C-terminal extension (R150, R154, K158, R162). Majority of the residues belonging to these potential binding epitopes are structurally distant from the binding groove and therefore contradict the view of the crystallographic structure. The PD conformation provides reasoning to why the C-terminal residues would influence the HA binding, as in this conformational state, the C-terminal tail is flexible enough to interact with the bound ligand [24, 73]. However, this conformational switch does not explain the role of residues, such as K38, that reside in the Link module outside the binding groove.

HABPs typically contain arginine-rich motifs vital to their ligand binding ability [67]. In the case of CD44, mapping all the potential HA binding arginines and other known binding residues onto the surface of HABD reveals an interaction surface so wide that it cannot be covered by a single HA oligomer. Due to such widespread nature of these residues, it has been suggested that CD44 could have multiple modes of HA binding that would cover different regions of the HABD surface [10].

Our work in *Publication I* focused on HABD–HA binding at the molecular level. The results reveal the existence of three mutually-exclusive binding modes, including the crystallographic pose analogous to the existing HA-bound X-ray structure of CD44 HABD [11]. Our subsequent analyses

focus on characterizing these binding modes, especially in the context of the previously-identified HA binding residues. Additionally, we show how the binding modes fit in with the multivalent binding of several HABDs to a single long HA polymer. Overall, our results create an atomistic picture of the CD44–HA binding dynamics and thereby expand the knowledge of this prototypic protein–carbohydrate interplay.

4.1 CD44 Binds Hyaluronan with Three Different Binding Modes

Our unbiased binding simulations (see Section 3.6) systematically showed an HA_{16} oligomer to bind CD44 HABD with three well-defined modes (Figure 4.1). The first mode of binding is identical to the crystallographic HABD–HA complex [11], and hence, referred to as the crystallographic binding mode. In this pose, HA occupies the primary binding groove while sharing a close interaction with the charged side-chain of R41. Further details of this well-documented binding complex can be obtained from Refs. [11, 72, 207, 208]. The second binding mode describes an interaction where the ligand lies on top of β1–α1 loop as well as β0, β8, and β9 sheets, lacking contacts with the hook-like β4–β5 loop, which is closely involved in the crystallographic binding mode. In this pose, the carbohydrate rings of HA lie parallel to the surface of HABD, and thus, we refer to it as the parallel mode. In the third observed binding mode, the HA oligomer occupies a region spanning from the β4–β5 loop to the C-terminus, such that it assumes a "vertical" orientation compared to the crystallographic binding. Hence, we refer to it as the upright binding mode.

In the first set of spontaneous binding simulations, where the HABD–HA distance was set to 1 nm, three out of four replicas finished in the parallel binding mode. In the remaining replica, the ligand bound to the crystallographic binding groove, but detached after ca. 1 μ*s* of simulation and formed the upright binding complex. In the second set of binding simulations, where the HABD–HA distance was set to ca. 4 nm, two out of five replicas finished in the parallel and two out of five in the upright binding mode. The fifth replica also displayed interaction between HA and the canonical binding residues, yet in a less defined manner. These observations show that the observed binding modes are stable at the microsecond timescales of the simulations. Furthermore, the parallel and upright modes appear to be more prone to form during the initial stages (*i.e.*, the first microsecond) of the CD44–HA interaction as compared to the crystallographic mode.

4.2 Characteristics of the Binding Modes

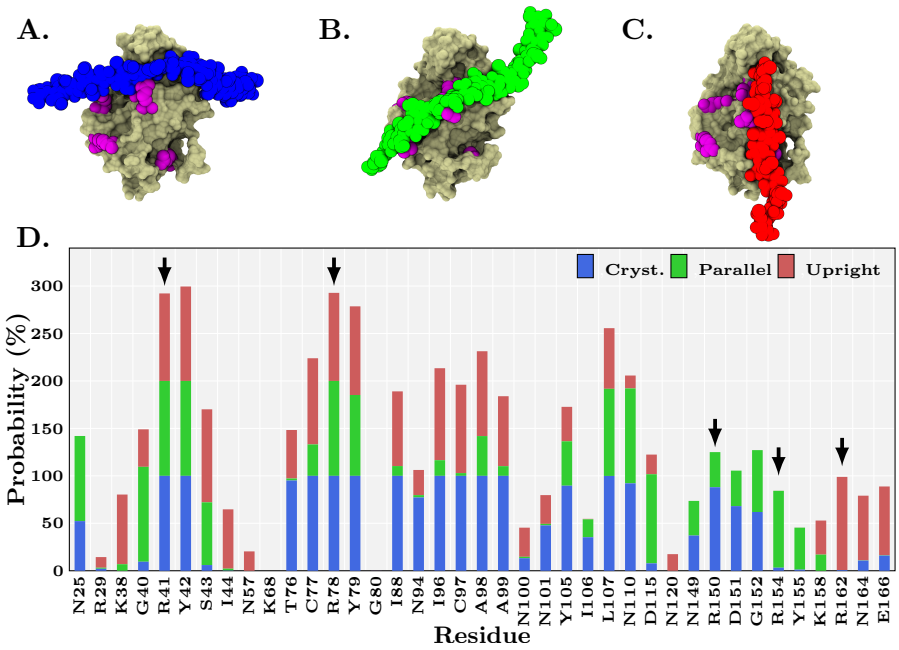

Figure 4.1: *Key results from* Publication I. *Snaphots of the A) crystallographic B) parallel, and C) upright binding modes. HA is color-coded to match the colors in panel D. D) HA binding probabilities of the experimentally-detected binding residues of CD44 HABD. The black arrows highlight the key arginine residues implicated in HA binding. The following references were used to gather the list of residues: [10, 11, 16, 21, 22, 76, 209].*

Having established the stability of the three binding modes, we set up simulations of each mode to further analyze them in the context of the previously-observed HA binding residues. Figure 4.1 lists these experimentally detected residues and shows how they are occupied by HA in our simulations. Strikingly, the combined fingerprint of all the binding modes matches that of the experiments, and thus, largely explains the observed spatial distribution of the residues.

There are two central HA-binding amino acids, R41 and R78, that are shared by all the binding modes in our simulations. Consistently, they are also among the most prominent residues whose mutation dramatically decreases HA recognition in the related wet-lab experiments [10,11,21,22,209]. Additionally, there are flanking residues more exclusive to a single binding mode. Mutating these flanking residues in previous experiments has been known to impair the binding, yet only as compound mutants [21]. This result could mean that the mutation of these secondary binding residues affects

the recognition of HA only when they collectively influence the formation of several binding modes.

Five experimentally-detected residues lack any interaction with HA in our simulations (Figure 4.1). These residues, however, reside on the "non-binding face" of HABD. Conversely, all the other known binding residues on the "binding face" of HABD share a significant probability to interact with the ligand in our simulations, at least in one binding mode. Taken together, these observations imply the existence of several HABD–HA binding poses.

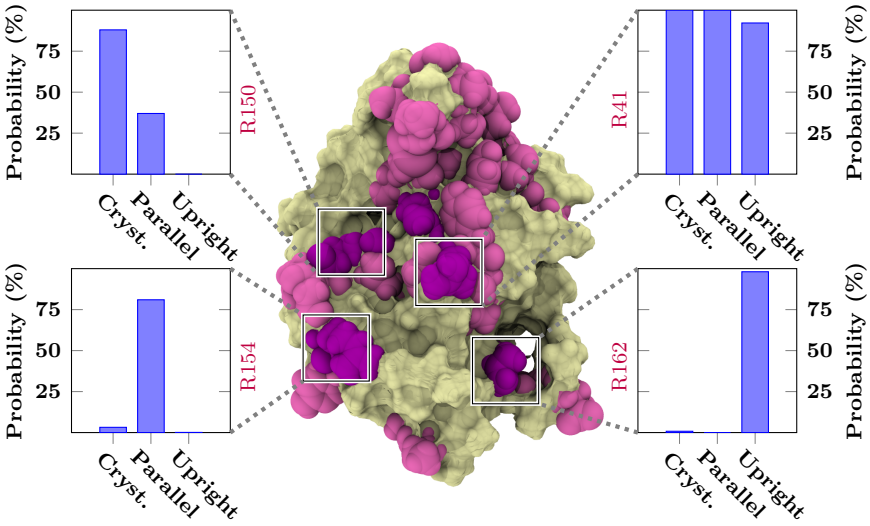

Figure 4.2: *Key results from Publication I. Experimentally-observed HA binding residues (see Figure 4.1) shown as pink beads on the surface of CD44 HABD. The key arginines are colored purple. The insets show the probability of each key arginine to interact with HA in each binding mode.*

To further characterize the binding modes, we focused on the arginine residues due to their role as prominent HA binders. Figure 4.2 illustrates four critical arginine residues and shows what was their contribution to each binding mode in our simulations. As stated earlier, R41 is the single most important HA binding residue of CD44 [10, 11, 21, 22, 209]. Providing rationale to its importance, our simulations show that R41 forms a basis of a central HA binding epitope shared by all the binding modes. That is, this region renders the highest number of contacts with HA across all the binding modes (see *Publication I* for details). Our data also show that each binding mode is further stabilized by a secondary arginine residue 3–4 nm away from the central epitope, forming a spatially divided R41-RX motif

where X depends on the binding mode such that $X_{crystallographic} = 150$, $X_{parallel} = 154$, and $X_{upright} = 162$. These findings indicate that CD44 recognizes HA oligomer with more than one binding epitope regardless of the binding mode.

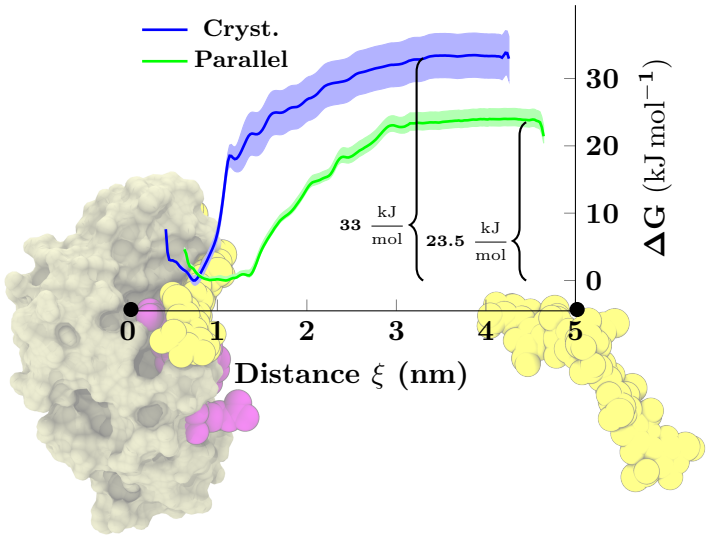

Figure 4.3: *Free energy profiles of HA_8 binding to CD44 HABD with the crystallographic (blue) and parallel (green) binding modes. The reaction coordinate (ξ) refers to the distance between HA and residues 75–79 of HABD (i.e., the "back" wall of the binding groove). The data are calculated with the umbrella sampling method [182]. The error estimates (shaded regions in the graphs) are standard errors obtained with the weighted histogram analysis method [184].*

To obtain more quantitative information on the strength of each binding mode, we measured their binding affinities using the umbrella sampling technique. Figure 4.3 shows that the crystallographic mode has a binding free energy of ca. -33 kJ mol^{-1} ($\sim$ 13.2 k_BT, $k_D = 2.75 \times 10^{-6}$ M), while the parallel mode has a lower value of -22 kJ mol^{-1} ($\sim$ 8.8 k_BT, $k_D = 1.96 \times 10^{-4}$ M). Strikingly, these simulation-derived values are in the same ballpark as the experimentally-measured ones in the literature (k_D = 13–27 μM) [11, 210]. Moreover, the clear difference of $\sim$ 4.4 k_BT in the measured affinities indicates that the crystallographic mode binds HA more strongly than the parallel mode. The fact that we observed multiple parallel mode complexes to form spontaneously in our simulations implies that it forms rather readily at the tested timescales and could thus represent a metastable state, which eventually leads to the formation of the crystallographic complex at longer

timescales. Due to their lower HA binding affinity, the weaker binding modes might also be used to bind other molecules. Namely, in addition to HA, CD44 is known to have other ligands, such as osteopontin and collagen [69].

4.3 CD44 Exhibits Spatially Restricted Motion Along Hyaluronan

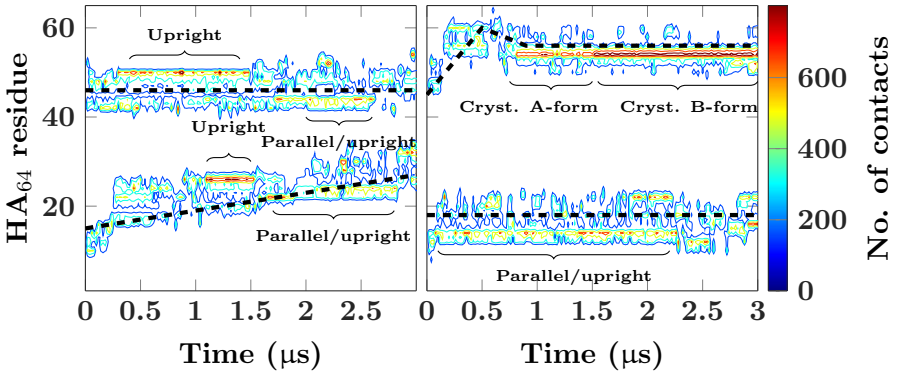

Figure 4.4: *Contour plots describing the contacts of HA residues with two CD44 proteins. The vertical axis lists the HA residues, the horizontal axis depicts the time into the simulation trajectory, and the contours show the number of HA–CD44 contacts according to the color bar on the right. The dashed lines guide the eye to see the motion of the two proteins along the HA strand. Data are shown for two independent replica systems.*

As discussed in Chapter 2, the size of HA polymers can vary greatly. Furthermore, the largest HA fragments have been observed to augment CD44 binding and clustering, suggesting that multivalent interactions play a key role in the recognition of HA [210, 211]. Learning more about such interactions is important because the clustering of CD44 is known to facilitate CD44-mediated signaling *via* the binding of accessory proteins [69, 77]. Hence, we explored the multivalent CD44–HA interactions by generating a system of two CD44 HABDs and a single HA_{64} polymer, which we then simulated with two three-microsecond simulation replicas. Figure 4.4 presents the main findings from these simulations as numbers of contacts between HABD and each saccharide residue of HA as a function of time.

In the first run, the other HABD (Figure 4.4, left) moves a stretch of ca. 20 saccharide residues relative to the HA chain. During this time, the observed HABD–HA complexes correspond primarily to the parallel and upright modes, which form spontaneously. These complexes last hundreds

of nanoseconds, after which they transiently (10–500 ns) dissociate, making the connection between HABD and HA looser. The relative motion occurs during these breaks. Although this motion is passive — *i.e.*, does not require external energy input — it might increase the probability of CD44s to cluster simply because it limits their degrees of freedom into one dimension (*i.e.*, parallel to the HA polymer). Moreover, because the motion is facilitated by the weaker binding modes, it could provide a physiological justification for the existence of such modes.

In the second run, we observe a spontaneous formation of the crystallographic complex between HABD and HA_{64} polymer (Figure 4.4, right). The binding occurs in two phases. First, HABD in the A-conformation binds to the HA polymer, placing the ligand into the crystallographic binding groove [11]. Second, after ca. 700 ns, the A-form conformation turns into the B-form complex, thus completing the binding by enabling more intimate contact between R41 side-chain and the bound HA. Both phases can be seen in Figure 4.4 as an increase in the measured contacts. The initial binding occurs at ca. 750 ns into the trajectory, after which the molecules stay firmly bound through the rest of the simulation, halting any relative motion of the protein along the HA_{64} polymer. These observations are well in line with another computational study stating that the B-form is the high-affinity conformation of R41 [72,73]. The observation also agree with our earlier affinity calculations, suggesting that the canonical crystallographic mode is indeed the primary CD44–HA binding mode.

4.4 Critical Assessment and Future Perspectives

Our results are consistent with previous experimental research. We are the first to record the steps of spontaneous CD44–HA binding with atomistic precision, thus confirming the current view of this process. This includes the A–B conformational transition leading to the formation of the crystallographic complex. Moreover, our residue-by-residue binding profile of the crystallographic complex — the only previously-known CD44–HA binding interaction — agrees with both experiments and simulations [11, 72]. Our results also provide novel insight into the recognition of HA, further explaining the previous experimental findings, such as the role of specific HA binding residues. What we neglected in our study, however, is the O–PD conformational transition seen with NMR [23, 24]. Another simulation study has shown that the increased flexibility of the PD conformation can also explain the role of certain C-terminal amino acids as HA binders [73]. However, while being a potential mechanism for regulating HA binding, the

PD conformation still does not explain all the observed binding residues. Furthermore, the idea of multiple binding modes is by no means mutually exclusive with the O–PD transition. That is, such mechanisms can coexist.

It can be problematic to sample the motion and binding of a relatively large and flexible ligand (HA_8) with umbrella sampling. The usefulness of the results depends on the choice of the reaction coordinate, as one cannot sample all the states of the system. We chose the distance between HA and the binding groove as our reaction coordinate because it is robust to measure with simulations. As a result, the binding process, starting from the initial (bound) and final (unbound) states, is well sampled, rendering the overall estimation of the binding affinity reliable. Further supporting this claim, our binding affinity values match those of the experimentally-derived dissociation constants, as noted above.

Our model of CD44 HABD lacked *N*-glycosylations, which are known to be important for the recognition of HA [14–17]. However, it is worthwhile to notice that the most relevant experimental studies have also used non-glycosylated CD44 constructs to draw their conclusions [10, 11, 209]. For example, many of the structural studies have used prokaryotic organisms — that lack the cellular machinery for complex *N*-glycosylations — to express their CD44 proteins. Our results from the non-glycosylated systems are therefore generally comparable with the experimental work. Moreover, this work without glycans sets the basis for further simulation studies to explore the role of the *N*-glycans in the recognition of HA.

4.5 Conclusions

In *Publication I*, we classified three binding modes of the CD44–HA interaction, termed as *crystallographic*, *parallel* and *upright*. The crystallographic mode refers to the canonical HABD–HA binding complex. Parallel and upright modes are two novel binding poses discovered through spontaneous binding in our unbiased simulations. Combining the residue-by-residue contact profiles of these three mutually-exclusive binding modes outputs a binding fingerprint that explains how the experimentally-observed, spatially widespread CD44 residues are all important in the recognition of HA. That is, the key HA binding residues, such as R41, are shared by all the binding modes, while the less important residues are exclusive to a single binding mode. We also recorded a spontaneous binding sequence between CD44 and HA and demonstrated how the weaker binding modes may be relevant in the motion of CD44 along a HA polymer. This relative motion may play a role in, *e.g.*, the clustering dynamics of CD44.

Chapter 5

N-glycans Regulate the CD44–Hyaluronan Interaction

HA recognition of CD44 is known to be regulated by the *N*-glycosylation of HABD [14–17]. Yet, previous studies have reported contrasting findings regarding the details of this regulation. While some glycan content seems to favor HA binding [76, 77], at least the presence of negatively-charged sialic acids generally interferes or inhibits it [14–17, 26, 75, 78]. This inhibition is intuitively explained by charge repulsion, as both sialic acids and HA bear a negative charge. However, given that all the currently available structural data of CD44–HA complexes are derived from non-glycosylated constructs [11, 23, 212, 212], the molecular mechanisms underlying the glycosylation-dependent regulation of HA binding have remained elusive.

Based on our previous study, we know that CD44 HABD contains various HA binding epitopes [191]. There is also evidence that the prominent effects of sialic acids as the prime regulators of HA binding could stem from the competitive inhibition of these HA binding epitopes. For instance, when neuraminidase was used to cleave the sialic acids from several melanoma cell lines with different levels of HA binding, the cells with the highest changes in HA binding levels released the least amount of sialic acids by weight [213]. This behavior indicates that the presence of sialic acids at specific sites is more important to HA binding than the overall number of sialic acids. Likewise, mutating each of the *N*-glycosylation sites individually in the inducible cell lines displayed low levels of binding with the N100A and N110A mutants but turned into completely active with the N25A and N120A mutants [16]. This observation implies that N25 and N120 could be the two most important sites of *N*-glycosylation regarding their ability to modulate HA binding. Supporting this notion, previous

computer simulations have shown sialylated *N*-glycans at N25 to occupy critical HA binding residues [27], such as R41, thus indirectly implying that the ability of these residues to bind HA is decreased in a glycosylated state of the protein.

In *Publication II*, we *in silico N*-glycosylated the CD44 HABD structure to uncover the role of the *N*-glycans in the regulation of HA binding [192]. To this end, we simulated *N*-glycosylated HABD without HA and with HA in an unbound state (*e.g.*, see Figure 5.1C–F). With these simulations, we show how complex *N*-glycans at N25, N100, and N110 cooperatively cover the canonical binding groove of CD44 HABD. That is, these glycans form a shield that significantly hinders the accessibility of the primary ligand binding site. The reduced accessibility of the crystallographic site, in turn, promotes the formation of the less shielded but lower affinity upright binding mode. This finding demonstrates a previously-unknown molecular mechanism to regulate the ligand-binding affinity of receptor proteins by promoting alternate binding sites through glycosylation.

In this work, we also deciphered how different glycoforms of CD44 HABD recognize HA. The results reveal that *N*-glycans hinder the binding of HA by preventing its entry to the binding site *via* both charge and steric effects. These observations provide a complete and realistic view into how different glycan profiles alter carbohydrate–protein interactions on the molecular level.

5.1 *N*-glycans Block the Canonical Hyaluronan Binding Site

To study the folding of the *N*-glycans on CD44, we constructed systems of glycosylated HABD lacking the HA ligand. The following simulations showed the ligand-binding surface of HABD to be consistently covered by the *N*-glycans (Table 5.1), especially in the fully glycosylated *myeloma* glycoforms. This covering effect was facilitated by the interlocking of the bulky glycans next to the crystallographic binding groove, thus effectively forming a glycan shield over the primary HA binding site. An inter-*N*-glycan contact map (Figure 5.1A) illustrates that the *N*-glycans at N25, N100, and N110 were the main contributors to this glycan shield. Such behavior was observed with both tested carbohydrate force fields, GLYCAM06 and CHARMM36, although the latter displayed shorter contacts between the *N*-glycans. In addition to the glycan shield, we observed stable electrostatic interactions and hydrogen bonds between several HA binding amino acids and N-glycans, especially between the central arginines (*i.e.*, R41 and R78)

and terminal sialic acids. Such interactions helped to maintain a stable occlusion of the crystallographic ligand-binding site. Overall, these observations imply that steric effects play an integral role in the glycosylation-dependent regulation of CD44–HA affinity.

Measuring the occupancy of HA binding residues by *N*-glycans revealed that the residues of the upright mode are ca. 55–65 % less covered by the glycans than those of the crystallographic and parallel modes (Table 5.1). This observation held for all the tested CD44 glycoforms. It was also consistent between the two simulation models, though the absolute values were ca. 15–25 % lower with the CHARMM36 model compared to the GLYCAM06 model. This finding implies that the relative propensity of the upright binding mode might be higher than that of the crystallographic mode in a glycosylated state of the receptor.

Table 5.1: *N-glycan occupancies in the residues of each binding mode. The contributions of each HABD residue to each binding mode are extracted from* Publication I *[191]. Data calculated with the AMBER ff99SB-ILDN + GLYCAM06 force field combination is shown without parentheses. Data calculated with CHARMM36 description is shown in parentheses. The numbers indicate how much of the HABD surface critical to hyaluronate binding is, on average, covered by N-glycans.*

Binding residues	***Myeloma asialo*: occupancy (%)**	***Myeloma monosialo*: occupancy (%)**
Cryst.	50 ± 13 (38 ± 8)	51 ± 8 (33 ± 7)
Parallel	48 ± 14 (36 ± 7)	46 ± 11 (36 ± 3)
Upright	29 ± 12 (24 ± 4)	30 ± 7 (20 ± 1)
	***Partial monosialo*: occupancy (%)**	***Full pentasaccharide*: occupancy (%)**
Cryst.	35 ± 18	32 ± 7
Parallel	34 ± 17	26 ± 6
Upright	21 ± 13	16 ± 4

Meanwhile, we know that the anti-CD44 antibody MEM-85 inhibits the binding of HA to a glycosylated HABD [22, 214]. To study these interactions further, our collaborators used $^{15}N/^{1}H$ HSQC NMR to probe the binding of both HA and MEM-85 to a non-glycosylated HABD. They calculated the HSQC spectra for free ^{15}N HABD; ^{15}N HABD with 3-fold molar excess of HA_6; ^{15}N HABD with 2-fold molar excess of MEM-85; and ^{15}N HABD with 3-fold molar excess of HA_6 and 2-fold molar excess of MEM-85. Interestingly, the resulting chemical shift perturbation spectra displayed coincidental binding of both HA_6 and MEM-85. The HA-induced changes

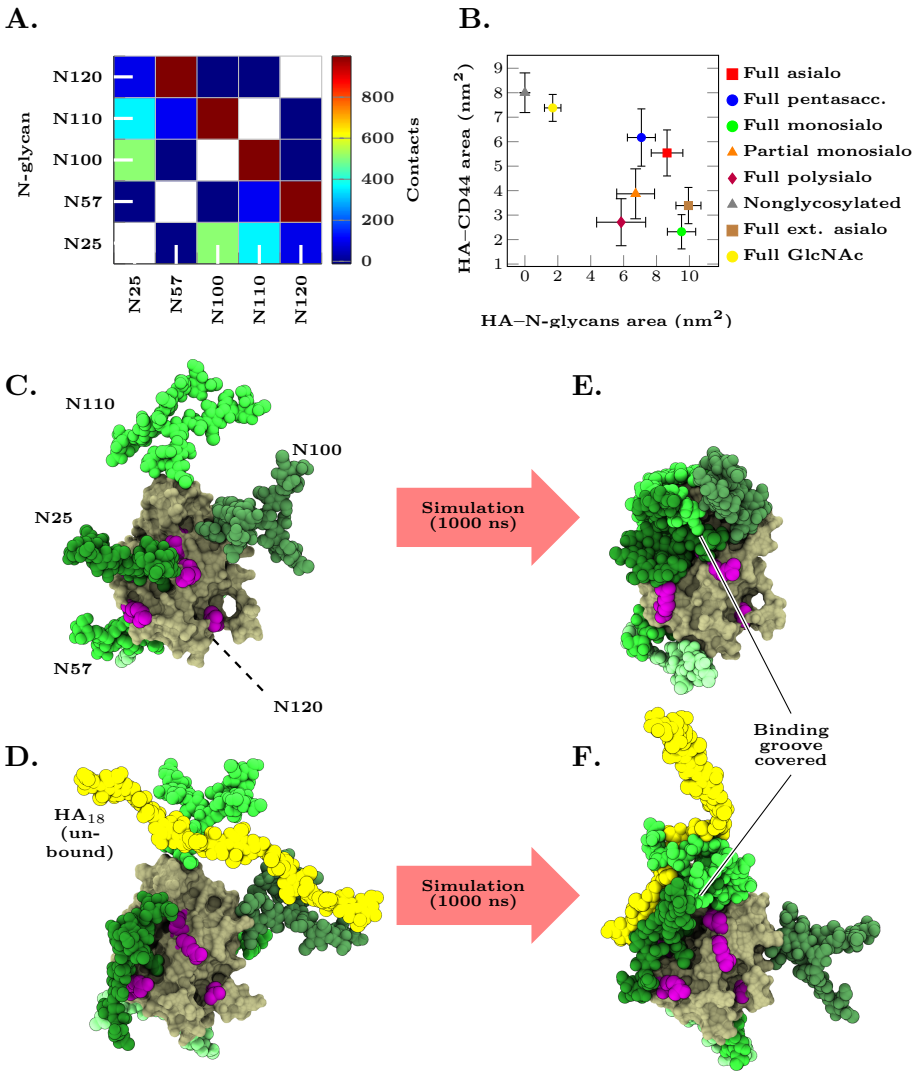

Figure 5.1: *Key results from Publication II. A) Number of contact matrix for the five N-glycans on CD44 HABD (*myeloma monosialo *glycoform). The data are averaged over both time and 15 replica simulations. B) Plot of CD44–HA interface area versus HA–N-glycans interface area in different glycoforms (see text and Figure 3.4). The data are calculated by the GROMACS tool* GMX SASA*. Error bars represent standard errors. C) and D) show starting structures of glycosylated HABD with (D) and without (C) HA. C) depicts the* myeloma monosialo *glycoform and D) corresponds to the* full monosialo *glycoform. E) and F) are corresponding snapshots after 1000 ns of simulation. It should be noted that these pictures represent only two example cases — more examples are available in the original publication.*

were localized to the R41-centered epitope as well as residues 75–105, which correspond to the crystallographic binding groove. The MEM-85-induced perturbations, on the other hand, were limited to the C-terminal portion of the receptor. These observations show that HA and MEM-85 have separate binding sites on a non-glycosylated CD44 HABD — *i.e.*, MEM-85 does not inhibit HA binding in such case.

In order for MEM-85 to inhibit HA binding to a glycosylated CD44, they must occupy the same binding site. In the case of MEM-85, we know that its binding site is centered around the C-terminal residues E160, Y161, and T163 [25]. As noted above, we have shown by MD simulations that these C-terminal residues are not perturbed by the attached *N*-glycans, suggesting that the binding of MEM-85 is not affected by the glycans. Hence, these facts point to the C-terminus being a binding epitope for both HA and MEM-85 in a highly glycosylated CD44 HABD. Supporting this idea, the C-terminal region forms an essential part of the upright binding mode [191]. Figure 5.2 also shows that the residues in this region are relatively more prone to bind HA compared to the central epitope residues R41 and R78 when the receptor is heavily glycosylated. Overall, these findings imply that *N*-glycosylation regulates CD44 by selectively promoting and inhibiting its HA binding sites.

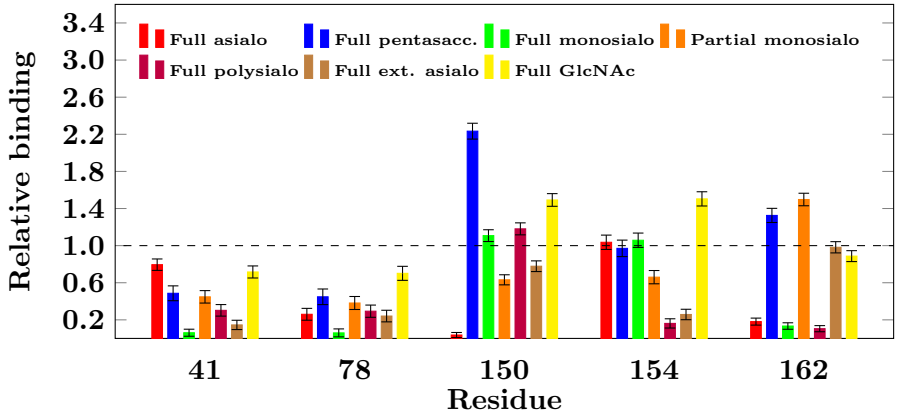

Figure 5.2: *Relative interaction probability of HA with five CD44 HABD arginine residues in different glycoforms. The data were calculated by counting the simulation frames in which a contact (see Chapter 3) is present. In this calculation, the first 200 ns of each simulation were excluded as an equilibration period. The obtained values were normalized with the reference values from the non-glycosylated system (*Publication I*) in order to obtain the relative interaction probabilities. As a result, values above 1 (dashed line) indicate an increase compared to the binding in the non-glycosylated case, while values below 1 indicate a decrease compared to the non-glycosylated case.*

5.2 Size and Charge of CD44 *N*-glycans Control the Binding of Hyaluronan

We then proceeded to explore the effect of different CD44 N-glycan profiles on HA binding. To this end, we generated systems in which HA_{18} and a glycosylated HABD are initially in an unbound state (see Section 3.6). This set-up ensured that the binding of the molecules during the simulations was spontaneous, *i.e.*, independent of the initial placement of the molecules.

Figure 5.1B shows the average final (t = 1000 ns) interface areas of HABD–HA and HA–N-glycans calculated from the spontaneous binding simulations. This analysis reveals how the glycan profile affects the binding of HA. Firstly, the non-sialylated glycoforms — *full GlcNAc*, *full pentasaccharide*, *full asialo*, and *full extended asialo* — display an inverse correlation between the HABD–HA interface area and size of the glycans, indicating steric effects or "glycan shielding" to be relevant in regulating the ligand binding. Secondly, the sialylated glycoforms show reduced binding, clearly implying that charge repulsion also plays a role in the recognition. The binding of HA was particularly hindered with the highly sialylated *full polysialo* glycoform. The *partial monosialo* glycoform displayed HABD–HA interface areas comparable to those of the *asialo* glycoforms. In the *partial monosialo* glycoform, the glycans at sites N100 and N110 were also less likely to fold over the canonical binding groove as they were unable to interact with the missing N25 glycan. Strikingly, the equally-sized *full monosialo* and *full extended asialo* glycoforms showed only minor differences in the measured interface areas, implying that a single sialic acid per glycan antenna might not be enough to induce extensive repulsion of the HA ligand.

Taken together, our findings indicate that the CD44 N-glycans control HA binding *via* two main mechanisms: steric blocking and charge repulsion. The steric blocking means that the largest N-glycans can act as a switch between high and low-affinity CD44–HA interaction by regulating the availability of the crystallographic ligand binding site. Accordingly, the smallest glycoforms may even strengthen the canonical HA binding by presenting additional interaction sites for the ligand. The charge repulsion component, on the other hand, refers to the number of sialic acids regulating the overall strength of the interaction so that a sufficiently high number of sialic acids can completely block the interaction.

5.3 Critical Assessment and Future Perspectives

The modeling of carbohydrates is a challenging task [7, 27, 215, 216]. This is largely due to the chemistry of these molecules posing unique challenges to the development of simulation force fields. For example, the typical properties of carbohydrates — such as, the high number of chiral centers, stereoisomerism and anomeric effect — result in an enormous state space, even when dealing with a single monosaccharide. Further complicating the matters, carbohydrates are often highly branched and attached to other biomolecules.

Despite these challenges, popular carbohydrate force fields, GLYCAM06 and CHARMM36, have been reported to reproduce some experimental values, such as intramolecular conformations, reasonably well [156, 217–219]. On the other hand, the GLYCAM06 model has also been shown to produce abnormal aggregation of β-D-glucose in solution when used with the TIP3P water model [220, 221]. Furthermore, GLYCAM06 has been reported to exhibit overtwisting of cellulose fibers in recent simulations [222]. Meanwhile, CHARMM36 is known to underestimate the hydration free energies of simple monosaccharides when used with the TIP3P water model [220]. These issues show that, like any force field, carbohydrate force fields are not perfect. They have to be used with caution, carefully assessing the plausibility of each step. In the publications of this Thesis, we mitigated the potential problems of the primary force field (GLYCAM06) by repeating the simulations with another force field (CHARMM36). The observed trends are similar with both models.

Providing further validation for our approach, our results are consistent with previous experimental findings. For example, our simulations show that different CD44 glycoforms affect HA binding in the same way — *i.e.*, regulated by the charge and size of the *N*-glycans — as in previous experiments in which the effect of CD44 glycan profiles on its liganding binding was quantified by affinity capillary electrophoresis [17]. Furthermore, other studies have also found glycosylations to influence the ligand binding properties of proteins, *e.g.*, by controlling their conformations [223], interfaces [224], or orientations [225].

Our findings are also compatible with previous computational work in which charge-pairing between terminal sialic acids and basic HA binding residues was determined to cause the inhibition of HA binding. Namely, we observed similar interactions in our simulations, indicating that charge-pairing plays a role in maintaining the occlusion of the essential binding residues.

We chose the CD44 glycoforms used in this Thesis to be as realistic as possible, given the current knowledge of these structures [65]. The complex type sialylated glycans mimicked the pathological phenotypes known to exist in cancer cells [14], while the glycoforms with shorter oligosaccharides modeled the phenotypes with higher affinity for HA [14, 17]. Supporting our choices, our simulations show qualitative agreement with experiments [14, 16, 17]. Still, to fully grasp the role of various glycan profiles for HA recognition, we require more information on the link between different glycoforms and HA binding states on various cell types. Fortunately, there have been recent advances in the development of targeted glycoproteomic methods for characterizing the glycosylation states of proteins [65, 226–228]. Likewise, the computational methods to study glycans through modeling and simulation have also developed rapidly over the recent decade [135, 136, 159, 160].

5.4 Conclusions

In *Publication II*, we *in silico* N-glycosylated CD44 HABD with several realistic N-glycan profiles to show how the attached oligosaccharides affect the binding of the HA ligand. The simulations revealed that the canonical crystallographic binding mode is readily shielded by large N-glycans due to the steric blocking of the crystallographic HA binding groove. The weaker upright mode binding site, on the other hand, remained relatively unaffected by the N-glycans, although the high concentration of sialic acid residues was enough to abrogate the binding in any case. Together with our NMR data, these findings suggested that a glycosylated CD44 binds HA predominantly with the weaker, but less hindered upright binding mode different from the canonical high-affinity interaction. This implies that the N-glycans can control which HA binding site is available on CD44, thus controlling its binding properties in a precise manner and affecting, *e.g.*, its clustering dynamics. On a larger scale, our results revealed a novel regulatory mechanism that uses glycosylations to alter receptor affinity by inhibiting specific ligand-binding sites while promoting others. These novel insights into protein–carbohydrate interactions might also be generalizable to other carbohydrate-binding proteins.

Chapter 6

Molecular Insights into the Activation of Janus Kinases

The molecular activation mechanism of cytokine receptor–JAK signaling has remained unclear [29]. The current view states that ligand binding induces conformational changes to pre-dimerized receptor subunits [229, 229–231]. Indeed, pre-dimerization has been observed for both homo and heterodimeric class I and II receptors, implying it to be a generic mechanism for this receptor family [28, 229]. Especially homodimeric class I receptor, such as EpoR and GHR, have been linked to this activation mechanism through biochemical and structural studies [28, 231, 232]. Yet, the details of the proposed conformational changes following ligand binding have remained enigmatic.

In *Publication III*, we proposed a novel paradigm for JAK activation through ligand-induced dimerization, which is short-circuited by hyperactivating oncogenic mutations like JAK2 V617F [194]. The experimental part of this work introduced a method of dual-colour single-molecule fluorescence imaging combined with posttranslational cell-surface labeling. This technique is able to quantify the monomer–dimer equilibrium of homodimeric class I cytokine receptors at the physiological concentration in the PMs of living cells. The analyses showed TpoR, EpoR and GHR to be monomeric and randomly distributed in an inactive state but effectively dimerize upon introducing their respective ligands. In addition, quantification of the monomer–dimer equilibrium between the ligand and JAK2 V617F-induced dimerization revealed that the mutation amplifies the dimerization and activation in an additive manner. These findings suggested that dimerization is the primary activation mechanism for receptor–JAK2 complexes and involves an intracellular JAK2 PK domain-mediated dimerization interface. However, the molecular details of this activation mech-

anism remained unclear.

The modeling and simulation part of the work (Author's contribution) entailed constructing and simulating a realistic 3D model of an activated cytokine receptor–JAK2 dimer embedded in a lipid bilayer [194]. The goal was to evaluate its stability and dynamics as both the wild-type and V617F mutant to reveal the atomic-level details unreachable by experimental means. We also used smaller models of the putative dimerization interfaces to measure the differences in affinity between the wild-type and mutant dimers. Here, our results of JAK activation agree with experimental data, strengthening the validity of the observations. We also revealed a novel membrane binding interaction of JAK2 that forms a prerequisite for the productive dimerization and activation of the receptor–JAK complexes.

6.1 JAK2 Signaling is Preceded by Dimerization

To reveal the mechanistic details of the dynamics of cytokine receptor–JAK2 dimers, we modeled two full-length JAK2s bound to either TpoR or EpoR in a pure POPC lipid bilayer. Both TpoR–JAK2 and EpoR–JAK2 systems showed the receptor complex to assume a stable upright orientation, with the long axis of the JAK2 FERM-SH2 domain aligning with the membrane normal, while the TK domains were residing the furthest from the lipids, as shown in Figure 6.1A. This orientation was maintained by a strong interaction between the FERM domain and the inner leaflet of the lipid bilayer. Integral to this interaction is the hydrophobic residue L224 as well as the positively charged patch of residues around the $\alpha 3$ helix in the F2 subdomain (Figure 6.1C). Notably, the interaction did not change when we altered the membrane composition from the pure POPC bilayer to a slightly more realistic POPC/cholesterol/phosphatidylinositol 4,5-bisphosphate (PIP2) (68/30/2 mol %) bilayer (Figure 6.1E). In fact, this change should only strengthen the membrane affinity of FERM as the positively charged patch in the F2 subdomain can interact with the negatively charged PIP2 lipids.

The coupling between the FERM domains and the lipids enforces an orientation of the receptor complex that, in turn, enables optimal inter-subunit interaction involving the putative PK domain-mediated dimerization interface [82]. Indeed, the dimerized PK domains maintained a stable interaction through the simulations, as shown in Figure 6.1D. The FERM-SH2 domains also displayed a relatively high number of inter-subunit contacts, but mainly due to their large surface area and proximity — *i.e.*, these domains did not form dimers in our simulations. The TK domains,

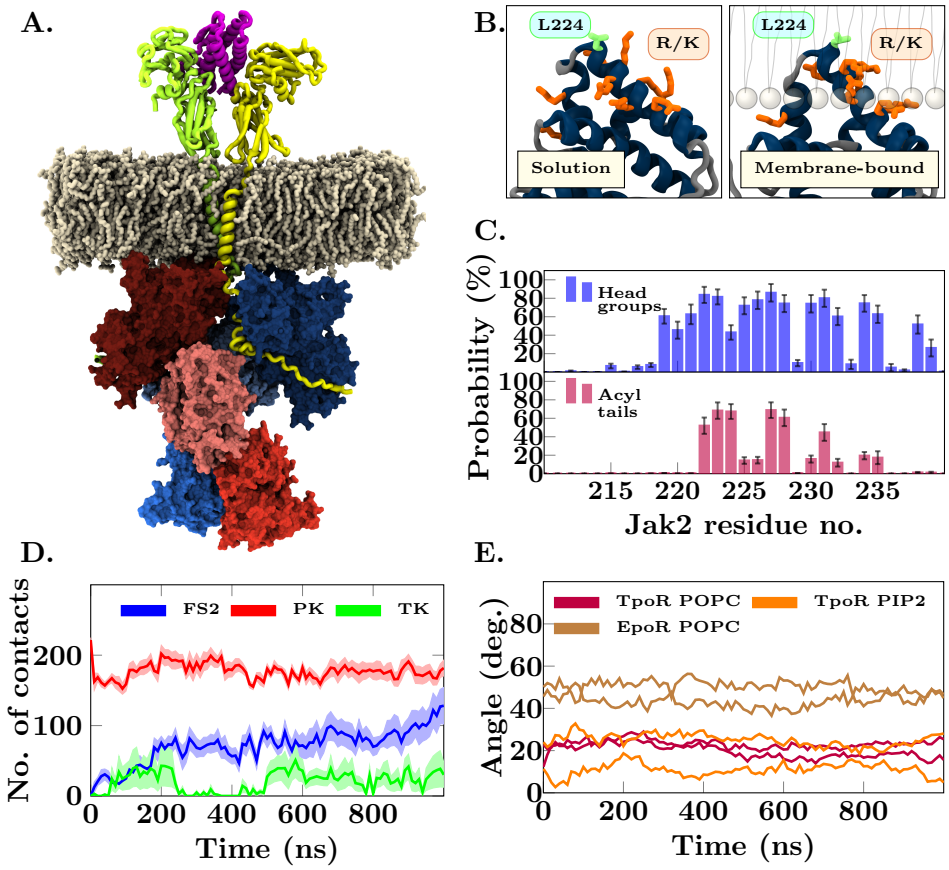

Figure 6.1: *Key results from Publication III. A) full-length JAK2–EpoR dimer model embedded in a pure POPC bilayer. The snapshot is taken at 1000 ns into a simulation trajectory. B) Membrane binding of the F2 subdomain of FERM. The side-chains of L224 (green) and the seven Lys and Arg residues (orange) in α3 change their orientation upon binding a membrane. C) Probability of residues in the F2 subdomain to interact (distance < 0.6 nm) with the lipid head groups (top) and the acyl chains (bottom). D) Domain–domain contacts between the JAK2 dimers. The PK domains (i.e., the intracellular dimerization interface) show stable interaction. E) Angle of the FERM domain — and thus the whole protein complex — when bound to TpoR in a POPC bilayer, mixed POPC/cholesterol/PIP2 bilayer, and when bound to EpoR in a pure POPC bilayer.*

initially modeled in a flexible conformation relative to the other domains, were found to be the most mobile parts of the signaling complex, sharing only transient inter-subunit contacts. This behavior supported the notion that dimerization — either through ligand binding or mutation — is driven by the inter-subunit PK–PK interface, which shifts the equilibrium away

from the intra-molecular PK–TK autoinhibitory interaction, thus freeing the TK domains to undergo *trans*-autophosphorylation [101]. This model is further supported by the fact that certain autoactivating mutations, such as R683S, are also located in the PK–TK autoinhibitory interface [30,233].

The recently-resolved X-ray structure of JAK2 FERM-SH2 domain in complex with the cytoplasmic tail of EpoR showed dimerization of the FERM domains through the juxtamembrane region of the bound EpoR [81]. The most prominent interactions in this EpoR-mediated dimer involved π stacking between the side-chains of EpoR W283 and JAK2 W298. Naturally, we also included them into our EpoR–JAK2 model that was constructed based on the X-ray structure [81]. However, these interactions dissociated during the simulations, and as a result, the juxtamembrane W283 residues were found to associate with the lipid bilayer (Figure 6.2). This behavior implies that the EpoR W283-mediated dimerization is relatively weak and that the associated residues are naturally drawn to the membrane-water interface.

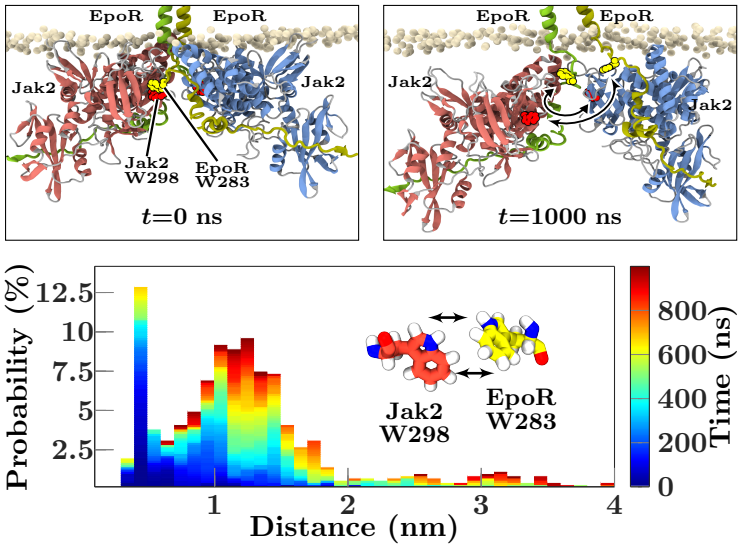

Figure 6.2: *Top panels: snapshots at the beginning (left) and end (right) of a JAK2-EpoR simulation. Yellow and red beads highlight EpoR W283 and JAK2 W298, respectively. The arrows on the right highlight the dissociation of EpoR W283 from the crystallographic dimerization pocket [81]. From JAK2, only the FERM domains are shown for clarity. Bottom panel: histogram of EpoR W283 and JAK2 W298 minimum distances. Color-coding indicates the simulation time. The initial state (at $t = 0$) corresponds to the highest probability peak seen at a distance of ca. 0.4 nm.*

6.2 Membrane-binding Controls JAK2 Activation

The most distinctive feature of the full JAK2–receptor model was the surprisingly stable interaction between the F2 subdomain of JAK2 FERM and the lipid bilayer. Residue L224 has a crucial role in this interaction, acting as an anchor that protrudes into the hydrophobic core of the bilayer. Highlighting its role in the membrane interaction, L224 is conserved in all JAK family members. Structurally, it is fully solvent-accessible in all known JAK structures and located in the C-terminal end of the α3-helix of FERM, which is surrounded by a patch of positively-charged residues presumed to mediate the membrane interaction, especially with charged lipids [234]. Upon membrane-binding, the L224-containing end of the α3-helix protrudes into the membrane core, leaving the charged residues on the level of lipid head groups, as shown in Figure 6.1B.

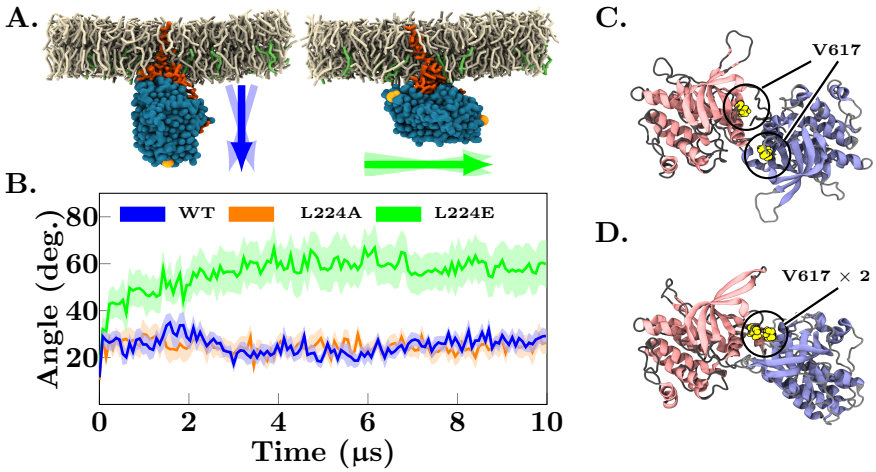

Figure 6.3: *A) The importance of L224 for the membrane binding of JAK2. The left panel shows a CG model of a wild-type JAK2 FERM-SH2 (blue) attached to TpoR (TM and IC domains, orange). The right panel shows the same construct having the L224E mutation. Arrows indicate the orientation of the FERM domain and its variation during the simulations. B) Angle of the FERM domain relative to the membrane normal in the tested JAK2 mutants. The measured vector was defined with the C_α atoms of residues 221 and 490. C) The AA starting structure of a wild-type JAK2 PK–PK dimer used in our simulations. D) Example of a final ($t = 1000$ ns) JAK2 PK–PK structure obtained from our simulations. The V617 residues (yellow beads) are highlighted with circles.*

To better evaluate the role of L224 in the membrane binding at microsecond timescales, we generated CG models of two JAK2 mutants, L224A and L224E, bound to a TpoR ΔECD monomer in a POPC/POPS bilayer and compared them to an analogous wild-type system. The resulting simu-

lations showed a striking difference in the orientation of the wild-type and L224E constructs: the wild-type remained in the stable, upright membrane binding pose, while the L224E assumed a less defined and more tilted orientation, as shown in Figure 6.3A–B. This difference shows that the single-residue membrane anchor is a crucial factor in orienting and stabilizing the entire protein complex.

When replicated in experiments, JAK2 L224E dramatically reduced the ligand-independent dimerization and activation of TpoR, EpoR, and GHR coupled to JAK2 V617F, suggesting that a proper membrane binding is required for JAK2-mediated dimerization [194]. Furthermore, FRAP experiments showed that the binding stability between JAK2 FERM-SH2 and TpoR at the PM of living cells was approximately 20-fold lower in case of the L224E mutant, implying that the orientation of the FERM-SH2 domain is essential for a productive receptor–JAK interaction. Further supporting this notion, visualization with TIRFM illustrated that JAK2 FERM-SH2 wild-type (fused to a green fluorescent protein) interacted more strongly with a surface micropatterned with either TpoR, EpoR, or GHR than the corresponding L224E mutant. All in all, these results confirmed that JAK2 L224E does not destabilize the conformation of JAK2, but affects the orientation of the protein complex, as predicted by the MD simulations.

6.3 Oncogenic Mutations Overstabilize the Dimer

To evaluate the differences between JAK2 wild-type and V617F, we simulated both of these forms as isolated PK–PK dimers (extracted from JAK1 PK–PK dimer; PDB:4l00; Figure 6.3C) and calculated their respective binding free energies using the MM/PBSA scheme [190]. These calculations yielded consistently lower free energy values for the V617F dimer relative to the wild-type with $\Delta\Delta G_{\mathrm{V617F}} = -32.7 \pm 16.2$ kJ/mol. This computationally-derived value agrees qualitatively with the experimentally measured binding free energy difference of $\Delta\Delta G_{\mathrm{V617F,exp}} = -5.2 \pm 1.1$ kJ/mol, calculated from the measured dimerization levels of receptor–JAK complexes [194]. The apparent quantitative discrepancy between the two values likely stems from the systematic underestimation of entropy by the MM/PBSA approach. Hence, these results imply that the V617F mutation indeed works by overstabilizing the cytosolic PK–PK dimerization interface.

Structurally, the JAK1-derived PK–PK interface involves the flexible SH2–PK linker segments, the αC helices, and the loops containing the V617 residues but in a conformation that lacks a direct V617–V617 interaction [82]. However, the unstructured nature of the SH2–PK linker

segments increased the conformational heterogeneity of the interface during the simulations, producing structures where a more direct inter-subunit V617–V617 contact was present (*e.g.*, see Figure 6.3D). This finding highlights the power of MD simulations in creating motion into static X-ray structures and reveals a potential model for the JAK2 PK–PK interaction close to that of the X-ray structure of JAK1.

6.4 Critical Assessment and Future Perspectives

The structures of the receptor–JAK dimers presented in this Thesis are models. That is, they might not represent native structures with 100 % accuracy. However, the models were built and verified based on various internal consistency checks. First, the coordinates of the individual domains were extracted from experimentally-resolved structures and then combined into a full structural model (see Section 3.6 and the original publication [194]). Second, the important interfaces, such as the PK–PK and FERM-SH2–receptor interactions, were taken from X-ray structures [81, 82]. Third, the inter-domain dynamics observed with MD simulations are similar to that observed previously with EM imaging of the homologous JAK1–gp130 holocomplex [196]. Fourth, the obtained simulation results agree with our experiments. So while the structure of our dimer complex is a model, it is well justified and passes all the internal consistency checks. In the future, techniques like cryo-EM are poised to eventually reveal the full structure of a homodimeric cytokine receptor–JAK complex, thus updating the view of its structural details. Meanwhile, our models represent the state-of-the-art understanding of this molecular interaction.

The CG simulations form a relevant part of the work in *Publication III*. As explained earlier, it is imperative to know the limitations of CG models to utilize them correctly. In this work, the CG simulations provided extra guidance when predicting the possible TpoR TM–TM dimer poses. Despite the well-documented issues with the used MARTINI 2 force field [169], these simulations were not done in isolation — *i.e.*, the obtained results were consistent with the outputs of other structure prediction tools (see Chapter 3) and can therefore be considered reliable. We also used a CG model when evaluating the role of JAK2 L224E mutant by changing the representative bead type from hydrophobic to charged. Therefore, the conclusions drawn from these systems are purely physical in nature, lacking any biological alterations, such as conformational changes, that would be out of the scope for CG models. This renders the observations solid and justifies the use of the CG model to speed up the sampling.

Despite the insight provided by *Publication III*, there are still questions regarding JAK activation that remain to be answered. For example, lipids are known to module the function of several membrane receptor proteins. However, very little is currently known about the link between membrane properties and JAK activation. In addition to lipids, the activation most probably involves considerable domain reorganization, such as the dissociation of autoinhibitory PK–TK complexes in *cis* as well as the association of the PK–PK domains in *trans*. Yet, the exact sequence of events leading to the activation remains to be elucidated. In the future, one could perform this task with either cryo-EM or long timescale MD simulations.

Although *Publication III* provides an explanation for JAK2 V617F and other related mutations at the PK–PK interface, there are a plethora of other activating mutants associated with JAKs or their cognate cytokine receptors [30]. Several mutations have, for example, been identified near the ATP-binding pocket of the kinase domains [30]. Another mutation hotspot resides in the TM domains of the cytokine receptors [235]. For instance, TpoR has been shown to harbor multiple clinically identified mutants in this region [235–237]. The exact mechanism of all these mutations and their relation to the activation mechanism remain elusive, although it is possible — and supported by the results of *Publication III* — that at least the majority of the TpoR TM mutations overstabilize the dimerization of the receptor complexes similar to JAK2 V617F.

6.5 Conclusions

Publication III provides novel insight into the activation of homodimeric cytokine receptors *via* JAK2. We modeled and simulated JAK2 both as individual domains and as full-length dimers together with its homodimeric cytokine receptor partners EpoR and TpoR. The simulations were integrally bridged to corresponding wet-lab experiments. Our data together with the data from the experiments clearly illustrated that the dimerization of two receptor–JAK2 subunits controls the activation of their kinase function. This dimerization can occur either via ligand binding or through autoactivating mutations, thus explaining the mechanism of operation of several clinically relevant JAK2 mutants, such as V617F. Finally, the joint view of our simulations and experiments suggests that a proper membrane anchoring of JAKs is a prerequisite for their receptor binding and dimerization. Overall, these findings increase our knowledge about the onset of blood cancers. They could also open up novel avenues for drug design against diseases, such as MPNs.

Chapter 7

Discussion

Today, computers are an invaluable tool in biophysical research. Simulation methods have evolved to not only complement experiments but also make predictions. Thanks to the ever-increasing computational power as well as the continuous development of novel algorithms, we can now see *in silico* how protein complexes behave at biologically relevant timescales. Moreover, this trend will continue in the future, which means that modeling and simulations will become increasingly important tools in understanding the emergence of diseases.

In this Thesis, modeling and MD simulations were used to study how structural modifications affect the regulation and activation mechanisms of two cell membrane-related proteins: CD44 and JAK2. Both proteins have a key role in human physiology, especially in cellular signaling. Due to their central role, both proteins are implicated in cancer — among other diseases — either through malfunctions, such as mutations, or abnormal expression profiles of proteins and glycans.

This Chapter places the work done in this Thesis into a broader context, speculating the type of added value it offers to the respective fields of both proteins. The Chapter also explores how each field is moving forward and highlights some of their key findings from the recent years. To this end, it places a special focus on outstanding medical issues and druggability.

7.1 Summary of Key Findings

In Publications I–II, we discovered that the interaction of the multifunctional cell-surface glycoprotein CD44 with its HA ligand might be more complex than previously thought. That is, in our simulations of non-glycosylated CD44, we found that the receptor binds HA with three dif-

ferent binding modes. The strongest of these modes is the canonical crystallographic interaction mode. The two novel binding modes — termed as parallel and upright modes — are likely to represent metastable states of the receptor–ligand interaction, yet they may nevertheless be physiologically relevant in processes, such as the clustering of CD44. Strikingly, our subsequent simulations showed that the upright binding mode could become a more dominant CD44–HA interaction mode in a glycosylated state of the receptor, as the crystallographic binding site is heavily covered by the N-glycans. This observation could mean that N-glycans can control receptor activity by promoting some binding sites over others, and raises a question of whether this is a common mechanism for many glycosylated receptor proteins.

In Publication III, we built a realistic model of an activated cytokine receptor–JAK2 dimer. Based on this model and experimental data, we then showed how clinically recognized mutations — most importantly the oncogenic JAK2 V617F — overstabilize the intracellular dimerization interface formed by the JAK2 pseudokinase domains. We also showed that the dimerization and activation rely on a proper membrane-anchoring of the receptor-JAK dimer. These results shed light into the emergence of myeloproliferative neoplasms — a diverse form of blood cancer whose onset is tightly associated with JAK2 mutations.

Taken together, modeling and simulations offer an appealing route to study nanoscale systems composed of biomolecules. For example, the soft and dynamic glycans would be relatively tedious to probe as precisely with other techniques. Similarly, piecing together realistic models of large protein complexes through modeling and simulations offers information that is currently not available in other ways. Therefore, the methodology in this Thesis is appropriate and serves significant benefits to the respective research projects. In the future, structural biology efforts will most likely yield more accurate information on both CD44 and JAK2, thereby corroborating or debunking the models proposed here. Given the fast development of novel software and force fields, it is also likely that future simulations will be able to generate complementary knowledge on these systems. However, for now, our models and results endow new prospects for understanding the medical issues related to both CD44 and JAK2.

7.2 CD44 in Future Cancer Therapeutics

Due to its central role in human physiology and pathology, the CD44–HA interaction has been studied extensively for the past three decades [69, 238,

239]. After a period of rigorous basic research, the focus of the studies is shifting towards a more applied science, aiming to exploit the CD44–HA interaction in the treatment of human disease, most notably cancer [240]. As both CD44 and HA are overexpressed in various tumor types, they provide an attractive target for selectivity between healthy and diseased tissues [69, 239, 241–243].

Currently, there are three main strategies of exploiting the CD44–HA interaction in the treatment of cancer [240]. In the first strategy, HA is used together with nanocarriers to deliver specific anti-cancer drugs directly to the tumor tissue expressing high numbers of CD44 [240, 244]. Such solutions typically involve the conjugation of HA to liposomes [245], polymeric nanoparticles [246], or development of HA-based, self-assembling nanocarriers [247, 248]. In the second strategy, HA is directly conjugated to an anti-cancer drug [249]. These type of solutions are actively pursued as they improve the bioavailability of drugs and reduce their toxicity. The third strategy involves targeting CD44 directly with humanized antibodies [250–253]. Indeed, several antibodies have been shown to block CD44 on cancer cells [240]. This blocking causes the cells to have a weaker interaction with the tumor matrix and, as a result, induces apoptosis [254].

A few promising CD44–HA-based treatments have reached the early stages of clinical trials. For example, ONCOFID^{TM}-P is a Paclitaxel–HA conjugate for the treatment of bladder cancer [249]. The conjugate enters tumor cells *via* CD44 internalization and localizes into lysosomes. The release of the drug in the lysosomes then results in a signaling cascade that leads to the apoptosis of cancerous cells. After phase I clinical trials showed a positive treatment outcome in 9 out of 15 patients, phase II studies were initiated with a larger cohort [255]. In addition to the ONCOFID^{TM}-P platform, antibodies have been designed to target CD44. U36 and BIWA (bivatuzumab) are two promising antibodies targeting the CD44v6 splice variant of CD44 [250–253]. Both drugs have reached phase I in clinical trials where they displayed significant antitumor potential in the treatment of head and neck squamous cell carcinoma. However, when bivatuzumab was conjugated to cytotoxic agent mertansine in another phase I study, severe skin toxicity was reported [256].

All these recent findings highlight the potential of CD44 in future cancer therapeutics. Yet, despite these developments, there are still blind spots in CD44 biology. Most notably, the full structure of the receptor is missing. Similarly, there is no high-resolution structure of CD44 with a realistic glycan profile — something where modeling and simulations can, however, offer indispensable benefits, as we have shown. This Thesis, as well as

earlier research, have already demonstrated how important *N*-glycans are for the ligand-binding function of CD44. Hence, their role in future therapeutics should not be neglected. Perhaps targeting both CD44 as well as specific tumor-related glycan profiles would in future offer benefits for the targeted delivery of anti-cancer drugs and antibodies. Moreover, models used in this Thesis could be expanded to include more components of the glycocalyx, thus generating an *in silico* platform to study drug permeability through the PM in a realistic and controlled manner.

7.3 Towards Disease-specific Janus Kinase Inhibitors

Recently, there have been multiple attempts to shed light into the structures of full-length JAK–receptor complexes [194, 257, 258], from which our study is a prime example [194]. The common goal is to acquire insight into the operational principles of these molecules through better structural understanding. Similar to our research, MD simulations have played an integral role in the recent modeling-based efforts [257, 258].

Parallel to our research, another study constructed a model of an active JAK2–EpoR dimer using experimental mutagenesis and X-ray data as a guide [257]. Their active dimer complex resembles that of ours in the overall arrangement of the domains, but they used different dimer interfaces for the PK–PK and TM–TM dimers that they discovered through MD simulations. In their structure, the clinically-relevant hotspot residues, including V617, interact with F3 subdomain of FERM in *cis*, thus forcing a suitable orientation for productive PK–PK interactions in *trans*. Similar to our research, these interactions would be overstabilized in the V617F mutant case. However, this view is different from our, more direct V617F–V617F contact model, and highlights the need for further studies.

Recently, a three-domain structure of human growth hormone receptor (HGR) was modeled based on data combined from small-angle X-ray scattering, NMR spectroscopy, X-ray crystallography, and MD simulations [258]. This model does not include JAKs but illustrates the dynamic nature of a single cytokine receptor subunit. Together with our study, these efforts highlight the growing trend of using modeling and MD simulations combined with other biophysical methods to uncover full-length structures of cytokine receptors and JAKs.

The final goal in this research is to design better treatments against disorders in JAK–STAT signaling [259, 260]. JAKs have posed a tempting target for drug developers ever since the discovery of the role of V617F in

oncogenesis [97–100]. Currently, some JAK inhibitors are already on the market, while many others are being developed [261]. Current inhibitors work by targeting the ATP-binding site of the TK domains, thus inhibiting their phosphate transfer properties. As a conserved and well-defined pocket, the ATP-binding site is relatively easy to target with drugs. There is also some evidence that ATP binding might stabilize the active conformation of the PK domain, which could in turn help to maintain the V617F-induced aberrant activation [93]. However, current JAK inhibitors do not provide effective selection between V617F and wild-type proteins [261].

The first JAK inhibitor to be approved by the Food and Drug Administration (FDA), as well as the European authorities, for the treatment of MPNs is ruxolitinib [262–264]. It is an orally effective drug that inhibits both JAK1 and JAK2 [264]. It has produced positive results and is thus approved for the treatment of myelofibrosis and polycythemia vera. Another JAK inhibitor, fedratinib, was also recently (2019) approved by the FDA for the treatment of these conditions [265]. Unlike ruxolitinib that binds the ATP-binding pocket, it is not known how fedratinib binds JAKs, yet it has shown selectivity towards JAK2 over the other JAK family members. Along with these drugs, there is a plethora of other compounds currently in clinical trials that show promises of treating MPNs effectively [266–269]. There are also several JAK inhibitors — either on the market or being tested — for the treatment of other diseases, such as rheumatoid arthritis [270, 271], inflammatory bowel disease [272], psoriasis [273], or atopic dermatitis [274].

A remarkable breakthrough in the field would be to develop a drug selective to the V617F mutant [20, 83]. Based on our research, this could be achieved by targeting the dimerization interface of the JAK2 PK domain instead of the ATP-binding pocket of the TK domain. This is because the PK–PK interface is central to the mutation-driven overactivation of JAK2. Such a strategy could optimistically lead to a decrease in JAK2 dimerization and thus a diminished overactivation of the JAK–STAT pathway. However, the dimerization interface of the PK domain might be challenging to target, as it lacks specific binding pockets like the ATP binding site. Another option arising from the results of our study could be to target the TM domains of cytokine receptors with fat-soluble drugs that partition into lipid membranes. Hindering the dimerization of the TM domains might shift the equilibrium towards inactive receptor–JAK monomers, thus reducing overactivation. Membrane-partitioning drugs could also be designed to interfere with the membrane binding of the FERM domain — again a feature that we have shown to affect the activation of JAK signaling.

It remains to be seen whether these strategies prove to be useful in treating diseases that are caused by malfunctioning JAKs. Nevertheless, our study has provided novel insight into the activation of JAKs, and thus, expanded these potential new avenues for treating JAK-driven disorders.

7.4 Concluding Remarks

This Thesis expanded the knowledge on two cell-surface proteins, CD44 and JAK2. Our models provided an atomic-scale understanding of these proteins and their ligands that can be expanded upon in future research. On a broader scale, we also showed how biomolecular modeling and simulations can help probe nanoscale phenomena that are difficult to access through other techniques. Similarly, this work highlighted how modeling and simulations can bring added value to biomolecular research when bridged to experiments. Such a combined setting allows a particular problem to be examined from several time and length scales, thus unlocking features that would remain hidden if investigated from only one perspective. Given that most future research problems require similar multidisciplinary problem solving, it is vital that researchers of all disciplines are aware of the available techniques — both computational and experimental — and can communicate with each other effectively.

References

[1] **Alberts, B., Bray, D., Hopkin, K., Johnson, A.D., Lewis, J., Raff, M., Roberts, K., and Walter, P.**, *Essential cell biology*. Garland Science, New York, USA. 2013.

[2] **Nelson, P.**, *Biological Physics*. W.H. Freeman and Company, New York, USA. 2004.

[3] **Nelson, D.L., Lehninger, A.L., and Cox, M.M.**, *Lehninger Principles of Biochemistry*. W.H. Freeman and Company, New York, USA. 2008.

[4] **Overington, J.P., Al-Lazikani, B., and Hopkins, A.L.**, How many drug targets are there? *Nature Reviews Drug Discovery*, 5(12): 993–996. 2006.

[5] **Corfield, A.P. and Berry, M.**, Glycan variation and evolution in the eukaryotes. *Trends in Biochemical Sciences*, 40(7): 351–359. 2015.

[6] **Gabius, H.J., Kaltner, H., Kopitz, J., and André, S.**, The glycobiology of the CD system: a dictionary for translating marker designations into glycan/lectin structure and function. *Trends in Biochemical Sciences*, 40(7): 360–376. 2015.

[7] **van Oosten, A.S. and Janmey, P.A.**, Extremely charged and incredibly soft: physical characterization of the pericellular matrix. *Biophysical Journal*, 104(5): 961. 2013.

[8] **Abraham, M.J., Murtola, T., Schulz, R., Páll, S., Smith, J.C., Hess, B., and Lindahl, E.**, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1: 19–25. 2015.

[9] **Aruffo, A., Stamenkovic, I., Melnick, M., Underhill, C.B., and Seed, B.**, CD44 is the principal cell surface receptor for hyaluronate. *Cell*, 61(7): 1303–1313. 1990.

[10] **Teriete, P., Banerji, S., Noble, M., Blundell, C.D., Wright, A.J., Pickford, A.R., Lowe, E., Mahoney, D.J., Tammi, M.I., and Kahmann, J.D.**, Structure of the regulatory hyaluronan binding domain in the inflammatory leukocyte homing receptor CD44. *Molecular Cell*, 13(4): 483–496. 2004.

[11] **Banerji, S., Wright, A.J., Noble, M., Mahoney, D.J., Campbell, I.D., Day, A.J., and Jackson, D.G.**, Structures of the Cd44–hyaluronan complex provide insight into a fundamental carbohydrate-protein interaction. *Nature Structural & Molecular Biology*, 14(3): 234–239. 2007.

[12] **Toole, B.P.**, Hyaluronan: from extracellular glue to pericellular cue. *Nature Reviews Cancer*, 4(7): 528–539. 2004.

[13] **Zöller, M.**, CD44: can a cancer-initiating cell profit from an abundantly expressed molecule? *Nature Reviews Cancer*, 11(4): 254–267. 2011.

[14] **Lesley, J., English, N., Perschl, A., Gregoroff, J., and Hyman, R.**, Variant cell lines selected for alterations in the function of the hyaluronan receptor CD44 show differences in glycosylation. *The Journal of Experimental Medicine*, 182(2): 431–437. 1995.

[15] **Katoh, S., Zheng, Z., Oritani, K., Shimozato, T., and Kincade, P.W.**, Glycosylation of CD44 negatively regulates its recognition of hyaluronan. *The Journal of Experimental Medicine*, 182(2): 419–429. 1995.

[16] **English, N.M., Lesley, J.F., and Hyman, R.**, Site-specific de-N-glycosylation of CD44 can activate hyaluronan binding, and CD44 activation states show distinct threshold densities for hyaluronan binding. *Cancer Research*, 58(16): 3736–3742. 1998.

[17] **Skelton, T.P., Zeng, C., Nocks, A., and Stamenkovic, I.**, Glycosylation provides both stimulatory and inhibitory effects on cell surface and soluble CD44 binding to hyaluronan. *The Journal of Cell Biology*, 140(2): 431–446. 1998.

[18] **Leonard, W.J. and O'Shea, J.J.**, Jaks and STATs: biological implications. *Annual Review of Immunology*, 16(1): 293–322. 1998.

[19] **O'Shea, J.J., Holland, S.M., and Staudt, L.M.**, JAKs and STATs in Immunity, Immunodeficiency, and Cancer. *New England Journal of Medicine*, 368(2): 161–170. 2013, pMID: 23301733.

[20] **O'Shea, J.J., Schwartz, D.M., Villarino, A.V., Gadina, M., McInnes, I.B., and Laurence, A.**, The JAK-STAT pathway: impact on human disease and therapeutic intervention. *Annual Review of Medicine*, 66: 311–328. 2015.

[21] **Peach, R.J., Hollenbaugh, D., Stamenkovic, I., and Aruffo, A.**, Identification of hyaluronic acid binding sites in the extracellular domain of CD44. *The Journal of Cell Biology*, 122(1): 257–264. 1993.

[22] **Bajorath, J., Greenfield, B., Munro, S.B., Day, A.J., and Aruffo, A.**, Identification of CD44 residues important for hyaluronan binding and delineation of the binding site. *Journal of Biological Chemistry*, 273(1): 338–343. 1998.

[23] **Takeda, M., Ogino, S., Umemoto, R., Sakakura, M., Kajiwara, M., Sugahara, K.N., Hayasaka, H., Miyasaka, M., Terasawa, H., and Shimada, I.**, Ligand-induced structural changes of the CD44 hyaluronan-binding domain revealed by NMR. *Journal of Biological Chemistry*, 281(52): 40089–40095. 2006.

[24] **Ogino, S., Nishida, N., Umemoto, R., Suzuki, M., Takeda, M., Terasawa, H., Kitayama, J., Matsumoto, M., Hayasaka, H., and Miyasaka, M.**, Two-state conformations in the hyaluronan-binding domain regulate CD44 adhesiveness under flow condition. *Structure*, 18(5): 649–656. 2010.

[25] **Škerlová, J., Král, V., Kachala, M., Fábry, M., Bumba, L., Svergun, D.I., Tošner, Z., Veverka, V., and Řezáčová, P.**, Molecular mechanism for the action of the anti-CD44 monoclonal antibody MEM-85. *Journal of Structural Biology*, 191(2): 214–223. 2015.

[26] **Katoh, S., Maeda, S., Fukuoka, H., Wada, T., Moriya, S., Mori, A., Yamaguchi, K., Senda, S., and Miyagi, T.**, A crucial role of sialidase Neu1 in hyaluronan receptor function of CD44 in T

helper type 2-mediated airway inflammation of murine acute asthmatic model. *Clinical & Experimental Immunology*, 161(2): 233–241. 2010.

[27] **Faller, C.E. and Guvench, O.**, Terminal sialic acids on CD44 N-glycans can block hyaluronan binding by forming competing intramolecular contacts with arginine sidechains. *Proteins: Structure, Function, and Bioinformatics*, 82(11): 3079–3089. 2014.

[28] **Atanasova, M. and Whitty, A.**, Understanding cytokine and growth factor receptor activation mechanisms. *Critical Reviews in Biochemistry and Molecular Biology*, 47(6): 502–530. 2012.

[29] **Hubbard, S.R.**, Mechanistic insights into regulation of JAK2 tyrosine kinase. *Frontiers in Endocrinology*, 8: 361. 2018.

[30] **Hammarén, H.M., Virtanen, A.T., Abraham, B.G., Peussa, H., Hubbard, S.R., and Silvennoinen, O.**, Janus kinase 2 activation mechanisms revealed by analysis of suppressing mutations. *Journal of Allergy and Clinical Immunology*, 143(4): 1549–1559. 2019.

[31] **Humphrey, W., Dalke, A., and Schulten, K.**, VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1): 33–38. 1996.

[32] **Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M.C., Tassani, S., Piva, F., Perez-amodio, S., Strippoli, P., and Canaider, S.**, An estimation of the number of cells in the human body. *Annals of Human Biology*, 40(6): 463–471. 2013.

[33] **Ridgway, N. and McLeod, R.**, *Biochemistry of Lipids, Lipoproteins and Membranes*. Elsevier, Amsterdam, Nertherlands. 2015.

[34] **Jones, R.A.L.**, *Soft Condensed Matter*. Oxford University Press, New York, USA. 2006.

[35] **Metcalf, D.**, Hematopoietic cytokines. *Blood*, 111(2): 485–491. 2008.

[36] **Belfiore, A. and LeRoith, D.**, *Principles of Endocrinology and Hormone Action*. Springer, New York, USA. 2018.

[37] **Valera, S., Hussy, N., Evans, R.J., Adami, N., North, R.A., Surprenant, A., and Buell, G.**, A new class of ligand-gated ion channel defined by P 2X receptor for extracellular ATP. *Nature*, 371(6497): 516–519. 1994.

[38] **Roeder, B.A., Kokini, K., Sturgis, J.E., Robinson, J.P., and Voytik-Harbin, S.L.**, Tensile mechanical properties of three-dimensional type I collagen extracellular matrices with varied microstructure. *Journal of Biomechanical Engineering*, 124(2): 214–222. 2002.

[39] **Schliwa, M. and Woehlke, G.**, Molecular motors. *Nature*, 422(6933): 759–765. 2003.

[40] **Grabon, A., Orłowski, A., Tripathi, A., Vuorio, J., Javanainen, M., Róg, T., Lönnfors, M., McDermott, M.I., Siebert, G., Somerharju, P., et al.**, Dynamics and energetics of the mammalian phosphatidylinositol transfer protein phospholipid exchange cycle. *Journal of Biological Chemistry*, 292(35): 14438–14455. 2017.

[41] **Lemmon, M.A. and Schlessinger, J.**, Cell signaling by receptor tyrosine kinases. *Cell*, 141(7): 1117–1134. 2010.

[42] **Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., and Erlich, H.A.**, Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839): 487–491. 1988.

[43] **Moremen, K.W., Tiemeyer, M., and Nairn, A.V.**, Vertebrate protein glycosylation: diversity, synthesis and function. *Nature Reviews Molecular Cell Biology*, 13(7): 448. 2012.

[44] **Stowell, S.R., Ju, T., and Cummings, R.D.**, Protein glycosylation in cancer. *Annual Review of Pathology: Mechanisms of Disease*, 10: 473–510. 2015.

[45] **Lee, H.S., Qi, Y., and Im, W.**, Effects of N-glycosylation on protein conformation and dynamics: Protein Data Bank analysis and molecular dynamics simulation study. *Scientific Reports*, 5(8926). 2015.

[46] **Rodgers, K. and McVey, M.**, Error-prone repair of DNA double-strand breaks. *Journal of Cellular Physiology*, 231(1): 15–24. 2016.

[47] **Kozmin, S., Slezak, G., Reynaud-Angelin, A., Elie, C., De Rycke, Y., Boiteux, S., and Sage, E.**, UVA radiation is highly mutagenic in cells that are unable to repair 7, 8-dihydro-8-oxoguanine in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 102(38): 13538–13543. 2005.

[48] **Wald, G.**, Defective color vision and its inheritance. *Proceedings of the National Academy of Sciences*, 55(6): 1347. 1966.

[49] **Diekmann, L., Pfeiffer, K., and Naim, H.Y.**, Congenital lactose intolerance is triggered by severe mutations on both alleles of the lactase gene. *BMC Gastroenterology*, 15(1): 36. 2015.

[50] **Goh, A.M., Coffill, C.R., and Lane, D.P.**, The role of mutant p53 in human cancer. *The Journal of Pathology*, 223(2): 116–126. 2011.

[51] **Dupuy, A.D. and Engelman, D.M.**, Protein area occupancy at the center of the red blood cell membrane. *Proceedings of the National Academy of Sciences*, 105(8): 2848–2852. 2008.

[52] **Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E.L.L.**, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3): 567–580. 2001.

[53] **Ahram, M., Litou, Z.I., Fang, R., and Al-Tawallbeh, G.**, Estimation of membrane proteins in the human proteome. *In Silico Biology*, 6(5): 379–386. 2006.

[54] **Almén, M.S., Nordström, K.J., Fredriksson, R., and Schiöth, H.B.**, Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biology*, 7(1): 1–14. 2009.

[55] **Ashcroft, F.M.**, From molecule to malady. *Nature*, 440(7083): 440–447. 2006.

[56] **Babon, J.J., Lucet, I.S., Murphy, J.M., Nicola, N.A., and Varghese, L.N.**, The molecular regulation of Janus kinase (JAK) activation. *Biochemical Journal*, 462(1): 1–13. 2014.

[57] **Hilger, D., Masureel, M., and Kobilka, B.K.**, Structure and dynamics of GPCR signaling complexes. *Nature Structural & Molecular Biology*, 25(1): 4. 2018.

[58] **Reitsma, S., Slaaf, D.W., Vink, H., Van Zandvoort, M.A., and oude Egbrink, M.G.**, The endothelial glycocalyx: composition, functions, and visualization. *Pflügers Archiv-European Journal of Physiology*, 454(3): 345–359. 2007.

[59] **Weinbaum, S., Tarbell, J.M., and Damiano, E.R.**, The structure and function of the endothelial glycocalyx layer. *Annual Reviews of Biomedical Engineering*, 9: 121–167. 2007.

[60] **Lipowsky, H.H.**, The endothelial glycocalyx as a barrier to leukocyte adhesion and its mediation by extracellular proteases. *Annals of Biomedical Engineering*, 40(4): 840–848. 2012.

[61] **Fu, B.M. and Tarbell, J.M.**, Mechano-sensing and transduction by endothelial surface glycocalyx: composition, structure, and function. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 5(3): 381–390. 2013.

[62] **Evanko, S.P., Tammi, M.I., Tammi, R.H., and Wight, T.N.**, Hyaluronan-dependent pericellular matrix. *Advanced Drug Delivery Reviews*, 59(13): 1351–1365. 2007.

[63] **Helenius, A. and Aebi, M.**, Intracellular functions of N-linked glycans. *Science*, 291(5512): 2364–2369. 2001.

[64] **Schauer, R.**, Sialic acids as regulators of molecular and cellular interactions. *Current Opinion in Structural Biology*, 19(5): 507–514. 2009.

[65] **Han, H., Stapels, M., Ying, W., Yu, Y., Tang, L., Jia, W., Chen, W., Zhang, Y., and Qian, X.**, Comprehensive characterization of the N-glycosylation status of CD44s by use of multiple mass spectrometry-based techniques. *Analytical and Bioanalytical Chemistry*, 404(2): 373–388. 2012.

[66] **Dahl, L., Dahl, I., Engström-Laurent, A., and Granath, K.**, Concentration and molecular weight of sodium hyaluronate in synovial fluid from patients with rheumatoid arthritis and other arthropathies. *Annals of the Rheumatic Diseases*, 44(12): 817–822. 1985.

[67] **Day, A.J. and Prestwich, G.D.**, Hyaluronan-binding proteins: tying up the giant. *Journal of Biological Chemistry*, 277(7): 4585–4588. 2002.

[68] **Tian, X., Azpurua, J., Hine, C., Vaidya, A., Myakishev-Rempel, M., Ablaeva, J., Mao, Z., Nevo, E., Gorbunova, V., and Seluanov, A.**, High-molecular-mass hyaluronan mediates the cancer resistance of the naked mole rat. *Nature*, 499(7458): 346–349. 2013.

[69] **Ponta, H., Sherman, L., and Herrlich, P.A.**, CD44: from adhesion molecules to signalling regulators. *Nature Reviews Molecular Cell Biology*, 4(1): 33–45. 2003.

[70] **Milner, C.M. and Day, A.J.**, TSG-6: a multifunctional protein associated with inflammation. *Journal of Cell Science*, 116(10): 1863–1873. 2003.

[71] **Banerji, S., Ni, J., Wang, S.X., Clasper, S., Su, J., Tammi, R., Jones, M., and Jackson, D.G.**, LYVE-1, a new homologue of the CD44 glycoprotein, is a lymph-specific receptor for hyaluronan. *The Journal of Cell Biology*, 144(4): 789–801. 1999.

[72] **Jamison II, F.W., Foster, T.J., Barker, J.A., Hills Jr, R.D., and Guvench, O.**, Mechanism of binding site conformational switching in the CD44–hyaluronan protein–carbohydrate binding interaction. *Journal of Molecular Biology*, 406(4): 631–647. 2011.

[73] **Favreau, A.J., Faller, C.E., and Guvench, O.**, CD44 receptor unfolding enhances binding by freeing basic amino acids to contact carbohydrate ligand. *Biophysical Journal*, 105(5): 1217–1226. 2013.

[74] **Rudy, W., Hofmann, M., Schwartz-Albiez, R., Zöller, M., Heider, K.H., Ponta, H., and Herrlich, P.**, The two major CD44 proteins expressed on a metastatic rat tumor cell line are derived from different splice variants: each one individually suffices to confer metastatic behavior. *Cancer Research*, 53(6): 1262–1268. 1993.

[75] **Zheng, Z., Cummings, R.D., Pummill, P.E., and Kincade, P.W.**, Growth as a solid tumor or reduced glucose concentrations in culture reversibly induce CD44-mediated hyaluronan recognition by Chinese hamster ovary cells. *Journal of Clinical Investigation*, 100(5): 1217. 1997.

[76] **Bartolazzi, A., Nocks, A., Aruffo, A., Spring, F., and Stamenkovic, I.**, Glycosylation of CD44 is implicated in CD44-mediated cell adhesion to hyaluronan. *The Journal of Cell Biology*, 132(6): 1199–1208. 1996.

[77] **Sleeman, J., Rudy, W., Hofmann, M., Moll, J., Herrlich, P., and Ponta, H.**, Regulated clustering of variant CD44 proteins increases their hyaluronate binding capacity. *The Journal of Cell Biology*, 135(4): 1139–1150. 1996.

[78] **Katoh, S., Miyagi, T., Taniguchi, H., Matsubara, Y.i., Kadota, J.i., Tominaga, A., Kincade, P.W., Matsukura, S., and Kohno, S.**, Cutting edge: an inducible sialidase regulates the hyaluronic acid binding ability of CD44-bearing human monocytes. *The Journal of Immunology*, 162(9): 5058–5061. 1999.

[79] **Paul, W.E. and Seder, R.A.**, Lymphocyte responses and cytokines. *Cell*, 76(2): 241–251. 1994.

[80] **Ward, A.C., Touw, I., and Yoshimura, A.**, The Jak-Stat pathway in normal and perturbed hematopoiesis. *Blood*, 95(1): 19–29. 2000.

[81] **Ferrao, R.D., Wallweber, H.J., and Lupardus, P.J.**, Receptor-mediated dimerization of JAK2 FERM domains is required for JAK2 activation. *eLife*, 7: e38089. 2018.

[82] **Toms, A.V., Deshpande, A., McNally, R., Jeong, Y., Rogers, J.M., Kim, C.U., Gruner, S.M., Ficarro, S.B., Marto, J.A., Sattler, M., Griffin, J.D., and Eck, M.J.**, Structure of a pseudokinase-domain switch that controls oncogenic activation of Jak kinases. *Nature Structural & Molecular Biology*, 20(10): 1221–1223. 2013.

[83] **Puleo, D.E., Kucera, K., Hammarén, H.M., Ungureanu, D., Newton, A.S., Silvennoinen, O., Jorgensen, W.L., and Schlessinger, J.**, Identification and characterization of JAK2 pseudokinase domain small molecule binders. *ACS medicinal Chemistry Letters*, 8(6): 618–621. 2017.

[84] **Lupardus, P.J., Ultsch, M., Wallweber, H., Kohli, P.B., Johnson, A.R., and Eigenbrot, C.**, Structure of the pseudokinase–kinase domains from protein kinase TYK2 reveals a mechanism for Janus kinase (JAK) autoinhibition. *Proceedings of the National Academy of Sciences*, 111(22): 8025–8030. 2014.

[85] **Zak, M., Hurley, C.A., Ward, S.I., Bergeron, P., Barrett, K., Balazs, M., Blair, W.S., Bull, R., Chakravarty, P., Chang, C., Crackett, P., Deshmukh, G., DeVoss, J., Dragovich, P.S., Eigenbrot, C., Ellwood, C., Gaines, S., Ghilardi, N., Gibbons, P., Gradl, S., Gribling, P., Hamman, C., Harstad, E., Hewitt, P., Johnson, A., Johnson, T., Kenny, J.R., Koehler, M.F.T., Kohli, P.B., Labadie, S., Lee, W.P., Liao, J., Liimatta, M., Mendonca, R., Narukulla, R., Pulk, R., Reeve,**

A., Savage, S., Shia, S., Steffek, M., Ubhayakar, S., van Abbema, A., Aliagas, I., Avitabile-Woo, B., Xiao, Y., Yang, J., and Kulagowski, J.J., Identification of C-2 hydroxyethyl imidazopyrrolopyridines as potent JAK1 inhibitors with favorable physicochemical properties and high selectivity over JAK2. *Journal of Medicinal Chemistry*, 56(11): 4764–4785. 2013.

[86] **Remy, I., Wilson, I.A., and Michnick, S.W.**, Erythropoietin receptor activation by a ligand-induced conformation change. *Science*, 283(5404): 990–993. 1999.

[87] **Moraga, I., Wernig, G., Wilmes, S., Gryshkova, V., Richter, C.P., Hong, W.J., Sinha, R., Guo, F., Fabionar, H., Wehrman, T.S., Krutzik, P., Demharter, S., Plo, I., Weissman, I.L., Minary, P., Majeti, R., Constantinescu, S.N., Piehler, J., and Garcia, K.C.**, Tuning cytokine receptor signaling by reorienting dimer geometry with surrogate ligands. *Cell*, 160(6): 1196–1208. 2015.

[88] **Wallweber, H.J., Tam, C., Franke, Y., Starovasnik, M.A., and Lupardus, P.J.**, Structural basis of recognition of interferon-α receptor by tyrosine kinase 2. *Nature Structural & Molecular Biology*, 21(5): 443. 2014.

[89] **Ferrao, R., Wallweber, H.J., Ho, H., Tam, C., Franke, Y., Quinn, J., and Lupardus, P.J.**, The structural basis for class II cytokine receptor recognition by JAK1. *Structure*, 24(6): 897–905. 2016.

[90] **McNally, R., Toms, A.V., and Eck, M.J.**, Crystal structure of the FERM-SH2 module of human Jak2. *PLoS One*, 11(5): e0156218. 2016.

[91] **Bandaranayake, R.M., Ungureanu, D., Shan, Y., Shaw, D.E., Silvennoinen, O., and Hubbard, S.R.**, Crystal structures of the JAK2 pseudokinase domain and the pathogenic mutant V617F. *Nature Structural & Molecular Biology*, 19(8): 754. 2012.

[92] **Andraos, R., Qian, Z., Bonenfant, D., Rubert, J., Vangrevelinghe, E., Scheufler, C., Marque, F., Régnier, C.H., De Pover, A., Ryckelynck, H., Bhagwat, N., Koppikar, P., Goel, A., Wyder, L., Tavares, G., Baffert, F., Pissot-Soldermann, C., Manley, P.W., Gaul, C., Voshol, H., Levine, R.L., Sellers, W.R., Hofmann, F., and Radimerski, T.**,

Modulation of activation-loop phosphorylation by JAK inhibitors is binding mode dependent. *Cancer Discovery*, 2(6): 512–523. 2012.

[93] **Hammarén, H.M., Ungureanu, D., Grisouard, J., Skoda, R.C., Hubbard, S.R., and Silvennoinen, O.**, ATP binding to the pseudokinase domain of JAK2 is critical for pathogenic activation. *Proceedings of the National Academy of Sciences*, 112(15): 4642–4647. 2015.

[94] **Feener, E.P., Rosario, F., Dunn, S.L., Stancheva, Z., and Myers, M.G.**, Tyrosine phosphorylation of Jak2 in the JH2 domain inhibits cytokine signaling. *Molecular and Cellular Biology*, 24(11): 4968–4978. 2004.

[95] **Hammarén, H.M., Virtanen, A.T., Raivola, J., and Silvennoinen, O.**, The regulation of JAKs in cytokine signaling and its breakdown in disease. *Cytokine*, 118: 48–63. 2019.

[96] **Dameshek, W.**, Some speculations on the myeloproliferative syndromes. *Blood*, 6(4): 372–375. 1951.

[97] **James, C., Ugo, V., Le Couédic, J.P., Staerk, J., Delhommeau, F., Lacout, C., Garcon, L., Raslova, H., Berger, R., Bennaceur-Griscelli, A., Villeval, J.L., Constantinescu, S.N., Casadevall, N., and Vainchenker, W.**, A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. *Nature*, 434(7037): 1144–1148. 2005.

[98] **Kralovics, R., Passamonti, F., Buser, A.S., Teo, S.S., Tiedt, R., Passweg, J.R., Tichelli, A., Cazzola, M., and Skoda, R.C.**, A gain-of-function mutation of JAK2 in myeloproliferative disorders. *New England Journal of Medicine*, 352(17): 1779–1790. 2005.

[99] **Levine, R.L., Wadleigh, M., Cools, J., Ebert, B.L., Wernig, G., Huntly, B.J., Boggon, T.J., Wlodarska, I., Clark, J.J., Moore, S., Adelsperger, J., Koo, S., Lee, J.C., Gabriel, S., Mercher, T., D'Andrea, A., Fröhling, S., Döhner, K., Marynen, P., Vandenberghe, P., Mesa, R.A., Tefferi, A., Griffin, J.D., Eck, M.J., Sellers, W.R., Meyerson, M., Golub, T.R., Lee, S.J., and Gilliland, D.G.**, Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell*, 7(4): 387–397. 2005.

[100] **Baxter, E.J., Scott, L.M., Campbell, P.J., East, C., Fourouclas, N., Swanton, S., Vassiliou, G.S., Bench, A.J., Boyd, E.M., Curtin, N., Scott, M.A., Erber, W.N., the Cancer Genome Project, and Green, A.R.**, Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. *The Lancet*, 365(9464): 1054–1061. 2005.

[101] **Silvennoinen, O. and Hubbard, S.R.**, Molecular insights into regulation of JAK2 in myeloproliferative neoplasms. *Blood*, 125(22): 3388–3392. 2015.

[102] **Zhao, L., Dong, H., Zhang, C.C., Kinch, L., Osawa, M., Iacovino, M., Grishin, N.V., Kyba, M., and Huang, L.J.s.**, A JAK2 interdomain linker relays Epo receptor engagement signals to kinase activation. *Journal of Biological Chemistry*, 284(39): 26988–26998. 2009.

[103] **Enkavi, G., Javanainen, M., Kulig, W., Róg, T., and Vattulainen, I.**, Multiscale simulations of biological membranes: the challenge to understand biological phenomena in a living substance. *Chemical Reviews*, 119(9): 5607–5774. 2019.

[104] **Zaccai, N.R., Serdyuk, I.N., and Zaccai, J.**, *Methods in Molecular Biophysics: Structure, Dynamics, Function for Biology and Medicine*. Cambridge University Press, Cambridge, UK. 2017.

[105] **Watson, J.D. and Crick, F.H.**, Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356): 737–738. 1953.

[106] **Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C., and Shore, V.**, Structure of myoglobin: A three-dimensional Fourier synthesis at 2 øA. resolution. *Nature*, 185(4711): 422–427. 1960.

[107] **White, S.H.**, The progress of membrane protein structure determination. *Protein Science*, 13(7): 1948–1949. 2004.

[108] **Dutta, S., Burkhardt, K., Young, J., Swaminathan, G.J., Matsuura, T., Henrick, K., Nakamura, H., and Berman, H.M.**, Data deposition and annotation at the worldwide protein data bank. *Molecular Biotechnology*, 42(1): 1–13. 2009.

[109] **Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E.**, The Protein Data Bank. *Nucleic Acids Research*, 28(1): 235–242. 2000.

[110] **Sikic, K., Tomic, S., and Carugo, O.**, Systematic comparison of crystal and NMR protein structures deposited in the protein data bank. *The Open Biochemistry Journal*, 4: 83. 2010.

[111] **Jayasinghe, S., Hristova, K., and White, S.H.**, MPtopo: A database of membrane protein topology. *Protein Science*, 10(2): 455–458. 2001.

[112] **Berman, H., Henrick, K., and Nakamura, H.**, Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, 10(12): 980–980. 2003.

[113] **Yu, S.M., McQuade, D.T., Quinn, M.A., Hackenberger, C.P., Gellman, S.H., Krebs, M.P., and Polans, A.S.**, An improved tripod amphiphile for membrane protein solubilization. *Protein Science*, 9(12): 2518–2527. 2000.

[114] **McGregor, C.L., Chen, L., Pomroy, N.C., Hwang, P., Go, S., Chakrabartty, A., and Privé, G.G.**, Lipopeptide detergents designed for the structural study of membrane proteins. *Nature Biotechnology*, 21(2): 171–176. 2003.

[115] **Shimizu, K., Cao, W., Saad, G., Shoji, M., and Terada, T.**, Comparative analysis of membrane protein structure databases. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1860(5): 1077–1091. 2018.

[116] **Sirohi, D., Chen, Z., Sun, L., Klose, T., Pierson, T.C., Rossmann, M.G., and Kuhn, R.J.**, The 3.8 Å resolution cryo-EM structure of Zika virus. *Science*, 352(6284): 467–470. 2016.

[117] **Míguez, A.S., Jiménez-Ortega, E., Ramírez-Escudero, M., Talens-Perales, D., Marín-Navarro, J., Polaina, J., Sanz-Aparicio, J., and Fernandez-Leiro, R.** *ACS Chemical Biology*, 15(1): 179–188. 2019.

[118] **Axelrod, D.**, Cell-substrate contacts illuminated by total internal reflection fluorescence. *Journal of Cell Biology*, 89(1): 141–145. 1981.

[119] **Axelrod, D., Koppel, D., Schlessinger, J., Elson, E., and Webb, W.W.**, Mobility measurement by analysis of fluorescence photobleaching recovery kinetics. *Biophysical Journal*, 16(9): 1055. 1976.

[120] **Chen, Y., Lagerholm, B.C., Yang, B., and Jacobson, K.**, Methods to measure the lateral diffusion of membrane lipids and proteins. *Methods*, 39(2): 147–153. 2006.

[121] **Yu, H.**, Extending the size limit of protein nuclear magnetic resonance. *Proceedings of the National Academy of Sciences*, 96(2): 332–334. 1999.

[122] **Wong-ekkabut, J. and Karttunen, M.**, The good, the bad and the user in soft matter simulations. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1858(10): 2529–2538. 2016.

[123] **Chothia, C. and Lesk, A.M.**, The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5(4): 823–826. 1986.

[124] **Martí-Renom, M.A., Stuart, A.C., Fiser, A., Sánchez, R., Melo, F., and Šali, A.**, Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29(1): 291–325. 2000.

[125] **Kaczanowski, S. and Zielenkiewicz, P.**, Why similar protein sequences encode similar three-dimensional structures? *Theoretical Chemistry Accounts*, 125(3-6): 643–650. 2010.

[126] **Šali, A. and Blundell, T.L.**, Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3): 779–815. 1993.

[127] **Aloy, P., Stark, A., Hadley, C., and Russell, R.B.**, Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins: Structure, Function, and Bioinformatics*, 53(S6): 436–456. 2003.

[128] **Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C.**, SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research*, 31(13): 3381–3385. 2003.

[129] **Fiser, A. and Ŝali, A.**, Modeller: generation and refinement of homology-based protein structure models. In *Methods in Enzymology*, volume 374, pp. 461–491, Elsevier. 2003.

[130] **Wallner, B. and Elofsson, A.**, All are not equal: a benchmark of different homology modeling programs. *Protein Science*, 14(5): 1315–1327. 2005.

[131] **Nayeem, A., Sitkoff, D., and Krystek Jr, S.**, A comparative study of available software for high-accuracy homology modeling: From sequence alignments to structural models. *Protein Science*, 15(4): 808–824. 2006.

[132] **Rossum, G.**, *Python Reference Manual*. CWI (Centre for Mathematics and Computer Science). 1995.

[133] **Lütteke, T., Bohne-Lang, A., Loss, A., Goetz, T., Frank, M., and von der Lieth, C.W.**, GLYCOSCIENCES. de: an Internet portal to support glycomics and glycobiology research. *Glycobiology*, 16(5): 71R–81R. 2006.

[134] **Park, S.J., Lee, J., Qi, Y., Kern, N.R., Lee, H.S., Jo, S., Joung, I., Joo, K., Lee, J., and Im, W.**, CHARMM-GUI Glycan Modeler for modeling and simulation of carbohydrates and glycoconjugates. *Glycobiology*, 29(4): 320–331. 2019.

[135] **Danne, R., Poojari, C., Martinez-Seara, H., Rissanen, S., Lolicato, F., Róg, T., and Vattulainen, I.**, doGlycans — tools for preparing carbohydrate structures for atomistic simulations of glycoproteins, glycolipids, and carbohydrate polymers for GROMACS. *Journal of Chemical Information and Modeling*, 57(10): 2401–2406. 2017.

[136] **Lemmin, T. and Soto, C.**, Glycosylator: a Python framework for the rapid modeling of glycans. *BMC Bioinformatics*, 20(1): 1–7. 2019.

[137] **Schlick, T.**, *Molecular Modeling and Simulation: an Interdisciplinary Guide*. Springer Science & Business Media, New York, USA. 2010.

[138] **Abraham, M., Van Der Spoel, D., Lindahl, E., and Hess, B.**, GROMACS user manual version 5.0.4. *Journal of Molecular Modeling*, 5: 1–298. 2014.

[139] **Wassenaar, T.A., Ingólfsson, H.I., Böckmann, R.A., Tieleman, D.P., and Marrink, S.J.**, Computational lipidomics with insane: a versatile tool for generating custom membranes for molecular simulations. *Journal of Chemical Theory and Computation*, 11(5): 2144–2155. 2015.

[140] **Jo, S., Kim, T., Iyer, V.G., and Im, W.**, CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 29(11): 1859–1865. 2008.

[141] **Wu, E.L., Cheng, X., Jo, S., Rui, H., Song, K.C., Dávila-Contreras, E.M., Qi, Y., Lee, J., Monje-Galvan, V., Venable, R.M., Klauda, J.B., and Im, W.**, CHARMM-GUI membrane builder toward realistic biological membrane simulations. *Journal of Computational Chemistry*, 35(27): 1997–2004. 2014.

[142] **Qi, Y., Ingólfsson, H.I., Cheng, X., Lee, J., Marrink, S.J., and Im, W.**, CHARMM-GUI Martini maker for coarse-grained simulations with the Martini force field. *Journal of Chemical Theory and Computation*, 11(9): 4486–4494. 2015.

[143] **Lee, J., Cheng, X., Swails, J.M., Yeom, M.S., Eastman, P.K., Lemkul, J.A., Wei, S., Buckner, J., Jeong, J.C., Qi, Y., Jo, S., Pande, V.S., Case, D.A., Brooks, C.L., MacKerell, A.D.J., Klauda, J.B., and Im, W.**, CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *Journal of Chemical Theory and Computation*, 12(1): 405–413. 2016.

[144] **Stansfeld, P.J., Goose, J.E., Caffrey, M., Carpenter, E.P., Parker, J.L., Newstead, S., and Sansom, M.S.**, MemProtMD: automated insertion of membrane protein structures into explicit lipid membranes. *Structure*, 23(7): 1350–1361. 2015.

[145] **Newport, T.D., Sansom, M.S.P., and Stansfeld, P.J.**, The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions. *Nucleic Acids Research*, 47(D1): D390–D397. 2019.

[146] **Rauscher, S., Gapsys, V., Gajda, M.J., Zweckstetter, M., de Groot, B.L., and Grubmüller, H.**, Structural ensembles of intrinsically disordered proteins depend strongly on force field: a

comparison to experiment. *Journal of Chemical Theory and Computation*, 11(11): 5513–5524. 2015.

[147] **Hess, B., Bekker, H., Berendsen, H.J.C., and Fraaije, J.G.E.M.**, LINCS: a linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12): 1463–1472. 1997.

[148] **Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A.**, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19): 5179–5197. 1995.

[149] **Darden, T., York, D., and Pedersen, L.**, Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12): 10089–10092. 1993.

[150] **Lyubartsev, A.P. and Rabinovich, A.L.**, Force field development for lipid membrane simulations. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1858(10): 2483–2497. 2016.

[151] **Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O., and Shaw, D.E.**, Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8): 1950–1958. 2010.

[152] **Best, R.B., Zhu, X., Shim, J., Lopes, P.E., Mittal, J., Feig, M., and MacKerell Jr, A.D.**, Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of Chemical Theory and Computation*, 8(9): 3257–3273. 2012.

[153] **Huang, J. and MacKerell Jr, A.D.**, CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34(25): 2135–2145. 2013.

[154] **Oostenbrink, C., Villa, A., Mark, A.E., and Van Gunsteren, W.F.**, A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry*, 25(13): 1656–1676. 2004.

[155] **Robertson, M.J., Tirado-Rives, J., and Jorgensen, W.L.**, Improved peptide and protein torsional energetics with the OPLS-AA force field. *Journal of Chemical Theory and Computation*, 11(7): 3499–3509. 2015.

[156] **Kirschner, K.N., Yongye, A.B., Tschampel, S.M., González-Outeiriño, J., Daniels, C.R., Foley, B.L., and Woods, R.J.**, GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *Journal of Computational Chemistry*, 29(4): 622–655. 2008.

[157] **Dickson, C.J., Madej, B.D., Skjevik, o.A., Betz, R.M., Teigen, K., Gould, I.R., and Walker, R.C.**, Lipid14: the amber lipid force field. *Journal of Chemical Theory and Computation*, 10(2): 865–879. 2014.

[158] **Klauda, J.B., Venable, R.M., Freites, J.A., O'Connor, J.W., Tobias, D.J., Mondragon-Ramirez, C., Vorobyov, I., MacKerell Jr, A.D., and Pastor, R.W.**, Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *The Journal of Physical Chemistry B*, 114(23): 7830–7843. 2010.

[159] **Mallajosyula, S.S., Guvench, O., and MacKerell, A.D.**, CHARMM Additive All-Atom Force Field for O-Glycan and N-Glycan Linkages in Carbohydrate-Protein Modeling. *Biophysical Journal*, 100(3): 526a. 2011.

[160] **Guvench, O., Mallajosyula, S.S., Raman, E.P., Hatcher, E., Vanommeslaeghe, K., Foster, T.J., Jamison, F.W., and MacKerell Jr, A.D.**, CHARMM additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate–protein modeling. *Journal of Chemical Theory and Computation*, 7(10): 3162–3180. 2011.

[161] **Marrink, S.J., De Vries, A.H., and Mark, A.E.**, Coarse grained model for semiquantitative lipid simulations. *The Journal of Physical Chemistry B*, 108(2): 750–760. 2004.

[162] **Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P., and De Vries, A.H.**, The MARTINI force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111(27): 7812–7824. 2007.

[163] **Monticelli, L., Kandasamy, S.K., Periole, X., Larson, R.G., Tieleman, D.P., and Marrink, S.J.**, The MARTINI coarse-grained force field: extension to proteins. *Journal of Chemical Theory and Computation*, 4(5): 819–834. 2008.

[164] **de Jong, D.H., Singh, G., Bennett, W.D., Arnarez, C., Wassenaar, T.A., Schäfer, L.V., Periole, X., Tieleman, D.P., and Marrink, S.J.**, Improved parameters for the martini coarse-grained protein force field. *Journal of Chemical Theory and Computation*, 9(1): 687–697. 2013.

[165] **Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., and Berendsen, H.J.**, GROMACS: fast, flexible, and free. *Journal of Computational Chemistry*, 26(16): 1701–1718. 2005.

[166] **Christen, M., Hünenberger, P.H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D.P., Heinz, T.N., Kastenholz, M.A., Kräutler, V., Oostenbrink, C., Peter, C., Trzesniak, D., and van Gunsteren, W.F.**, The GROMOS software for biomolecular simulation: GROMOS05. *Journal of Computational Chemistry*, 26(16): 1719–1751. 2005.

[167] **Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L., and Schulten, K.**, Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16): 1781–1802. 2005.

[168] **Riniker, S., Allison, J.R., and van Gunsteren, W.F.**, On developing coarse-grained models for biomolecular simulation: a review. *Physical Chemistry Chemical Physics*, 14(36): 12423–12430. 2012.

[169] **Javanainen, M., Martinez-Seara, H., and Vattulainen, I.**, Excessive aggregation of membrane proteins in the Martini model. *PLoS One*, 12(11): e0187936. 2017.

[170] **Wassenaar, T.A., Pluhackova, K., Böckmann, R.A., Marrink, S.J., and Tieleman, D.P.**, Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. *Journal of Chemical Theory and Computation*, 10(2): 676–690. 2014.

[171] **Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shan, Y., and Wriggers, W.**, Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002): 341–346. 2010.

[172] **Hockney, R.W., Goel, S., and Eastwood, J.**, Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, 14(2): 148–158. 1974.

[173] **Van Gunsteren, W.F. and Berendsen, H.J.**, A leap-frog algorithm for stochastic dynamics. *Molecular Simulation*, 1(3): 173–185. 1988.

[174] **Beberg, A.L., Ensign, D.L., Jayachandran, G., Khaliq, S., and Pande, V.S.**, Folding@ home: Lessons from eight years of volunteer distributed computing. In *2009 IEEE International Symposium on Parallel & Distributed Processing*, pp. 1–8, IEEE. 2009.

[175] **Kutzner, C., Páll, S., Fechner, M., Esztermann, A., de Groot, B.L., and Grubmüller, H.**, Best bang for your buck: GPU nodes for GROMACS biomolecular simulations. *Journal of Computational Chemistry*, 36(26): 1990–2008. 2015.

[176] **Berendsen, H.J., Postma, J.v., van Gunsteren, W.F., DiNola, A., and Haak, J.R.**, Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8): 3684–3690. 1984.

[177] **Evans, D.J. and Holian, B.L.**, The Nose–Hoover thermostat. *The Journal of Chemical Physics*, 83(8): 4069–4074. 1985.

[178] **Parrinello, M. and Rahman, A.**, Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12): 7182–7190. 1981.

[179] **Noether, E.**, Invariante Variationsprobleme, Nachrichten von der Königlichen. *Gesellschaft der Wissenschaften zu Göttingen*, pp. 234–257. 1918.

[180] **Kästner, J.**, Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6): 932–942. 2011.

[181] **Roux, B.**, The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1-3): 275–282. 1995.

[182] **Torrie, G.M. and Valleau, J.P.**, Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2): 187–199. 1977.

[183] **Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H., and Kollman, P.A.**, The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8): 1011–1021. 1992.

[184] **Hub, J.S., De Groot, B.L., and Van Der Spoel, D.**, g_wham — A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *Journal of Chemical Theory and Computation*, 6(12): 3713–3720. 2010.

[185] **Genheden, S. and Ryde, U.**, The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, 10(5): 449–461. 2015.

[186] **Gray, J.J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A., and Baker, D.**, Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, 331(1): 281–299. 2003.

[187] **Hou, T., Wang, J., Li, Y., and Wang, W.**, Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 51(1): 69–82. 2011.

[188] **Wang, C., Nguyen, P.H., Pham, K., Huynh, D., Le, T.B.N., Wang, H., Ren, P., and Luo, R.**, Calculating protein–ligand binding affinities with MMPBSA: Method and error analysis. *Journal of Computational Chemistry*, 37(27): 2436–2446. 2016.

[189] **Baker, N.A., Sept, D., Joseph, S., Holst, M.J., and McCammon, J.A.**, Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18): 10037–10041. 2001.

[190] **Kumari, R., Kumar, R., Consortium, O.S.D.D., and Lynn, A.**, g_mmpbsa — A GROMACS tool for high-throughput MM-PBSA calculations. *Journal of Chemical Information and Modeling*, 54(7): 1951–1962. 2014.

[191] **Vuorio, J., Vattulainen, I., and Martinez-Seara, H.**, Atomistic fingerprint of hyaluronan–CD44 binding. *PLoS Computational Biology*, 13(7): e1005663. 2017.

[192] **Vuorio, J., Škerlová, J., Fábry, M., Veverka, V., Vattulainen, I., Řezáčová, P., and Martinez-Seara, H.**, N-Glycosylation can selectively block or foster different receptor–ligand binding modes. *Scientific Reports*, 11(5239). 2021.

[193] **Varki, A., Cummings, R.D., Aebi, M., Packer, N.H., Seeberger, P.H., Esko, J.D., Stanley, P., Hart, G., Darvill, A., Kinoshita, T., Prestegard, J.J., Schnaar, R.L., Freeze, H.H., Marth, J.D., Bertozzi, C.R., Etzler, M.E., Frank, M., Vliegenthart, J.F., Lütteke, T., Perez, S., Bolton, E., Rudd, P., Paulson, J., Kanehisa, M., Toukach, P., Aoki-Kinoshita, K.F., Dell, A., Narimatsu, H., York, W., Taniguchi, N., and Kornfeld, S.**, Symbol nomenclature for graphical representations of glycans. *Glycobiology*, 25(12): 1323–1324. 2015.

[194] **Wilmes, S., Hafer, M., Vuorio, J., Tucker, J.A., Winkelmann, H., Löchte, S., Stanly, T.A., Prieto, K.D.P., Poojari, C., Sharma, V., Richter, C., Kurre, R., Hubbard, S., Garcia, C., Moraga, I., Vattulainen, I., Hitchcock, I.S., and Piehler, J.**, Mechanism of homodimeric cytokine receptor activation and dysregulation by oncogenic mutations. *Science*, 367(6478): 643–652. 2020.

[195] **Polyansky, A.A., Chugunov, A.O., Volynsky, P.E., Krylov, N.A., Nolde, D.E., and Efremov, R.G.**, PREDDIMER: a web server for prediction of transmembrane helical dimers. *Bioinformatics*, 30(6): 889–890. 2014.

[196] **Lupardus, P.J., Skiniotis, G., Rice, A.J., Thomas, C., Fischer, S., Walz, T., and Garcia, K.C.**, Structural snapshots of full-length Jak1, a transmembrane gp130/IL-6/IL-6Rα cytokine receptor complex, and the receptor-Jak1 holocomplex. *Structure*, 19(1): 45–55. 2011.

[197] **Syed, R.S., Reid, S.W., Li, C., Cheetham, J.C., Aoki, K.H., Liu, B., Zhan, H., Osslund, T.D., Chirino, A.J., Zhang, J., Finer-Moore, J., Elliott, S., Sitney, K., Katz, B.A., Matthews, D.J., Wendoloski, J.J., Egrie, J., and Stroud, R.M.**, Efficiency of signalling through cytokine receptors depends critically on receptor orientation. *Nature*, 395(6701): 511–516. 1998.

[198] **Li, Q., Wong, Y.L., Huang, Q., and Kang, C.**, Structural insight into the transmembrane domain and the juxtamembrane region of the erythropoietin receptor in micelles. *Biophysical Journal*, 107(10): 2325–2336. 2014.

[199] **Jones, A.V., Kreil, S., Zoi, K., Waghorn, K., Curtis, C., Zhang, L., Score, J., Seear, R., Chase, A.J., Grand, F.H., White, H., Zoi, C., Loukopoulos, D., Terpos, E., Vervessou, E.C., Schultheis, B., Emig, M., Ernst, T., Lengfelder, E., Hehlmann, R., Hochhaus, A., Oscier, D., Silver, R.T., Reiter, A., and Cross, N.C.P.**, Widespread occurrence of the JAK2 V617F mutation in chronic myeloproliferative disorders. *Blood*, 106(6): 2162–2168. 2005.

[200] **Campbell, P.J., Griesshammer, M., Dohner, K., Dohner, H., Kusec, R., Hasselbalch, H.C., Larsen, T.S., Pallisgaard, N., Giraudier, S., Le Bousse-Kerdilés, M.C., Desterke, C., Guerton, B., Dupriez, B., Bordessoule, D., Fenaux, P., Kiladjian, J., Viallard, J.F., Briére, J., Harrison, C.N., Green, A.R., and Reilly, J.T.**, V617F mutation in JAK2 is associated with poorer survival in idiopathic myelofibrosis. *Blood*, 107(5): 2098–2100. 2006.

[201] **Zaleskas, V.M., Krause, D.S., Lazarides, K., Patel, N., Hu, Y., Li, S., and Van Etten, R.A.**, Molecular pathogenesis and therapy of polycythemia induced in mice by JAK2 V617F. *PloS One*, 1(1): e18. 2006.

[202] **Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L.**, Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2): 926–935. 1983.

[203] **UniProt Consortium**, UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1): D158–D169. 2017.

[204] **UniProt Consortium**, UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1): D506–D515. 2019.

[205] **Pronk, S., Páll, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M.R., Smith, J.C., Kasson, P.M., van der Spoel, D., Hess, B., and Lindahl, E.**, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, p. btt055. 2013.

[206] **Bussi, G., Donadio, D., and Parrinello, M.**, Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1): 4101. 2007.

[207] **Jana, M. and Bandyopadhyay, S.**, Restricted dynamics of water around a protein–carbohydrate complex: Computer simulation studies. *The Journal of Chemical Physics*, 137(5): 055102. 2012.

[208] **Jana, M. and Bandyopadhyay, S.**, Conformational flexibility of a protein–carbohydrate complex and the structure and ordering of surrounding water. *Physical Chemistry Chemical Physics*, 14(18): 6628–6638. 2012.

[209] **Takeda, M., Terasawa, H., Sakakura, M., Yamaguchi, Y., Kajiwara, M., Kawashima, H., Miyasaka, M., and Shimada, I.**, Hyaluronan recognition mode of CD44 revealed by cross-saturation and chemical shift perturbation experiments. *Journal of Biological Chemistry*, 278(44): 43550–43555. 2003.

[210] **Wolny, P.M., Banerji, S., Gounou, C., Brisson, A.R., Day, A.J., Jackson, D.G., and Richter, R.P.**, Analysis of CD44-hyaluronan interactions in an artificial membrane system insights into the distinct binding properties of high and low molecular weight hyaluronan. *Journal of Biological Chemistry*, 285(39): 30170–30180. 2010.

[211] **Yang, C., Cao, M., Liu, H., He, Y., Xu, J., Du, Y., Liu, Y., Wang, W., Cui, L., Hu, J., et al.**, The high and low molecular weight forms of hyaluronan have distinct effects on CD44 clustering. *Journal of Biological Chemistry*, 287(51): 43094–43107. 2012.

[212] **Liu, L.K. and Finzel, B.C.**, Fragment-Based Identification of an Inducible Binding Site on Cell Surface Receptor CD44 for the Design of Protein–Carbohydrate Interaction Inhibitors. *Journal of Medicinal Chemistry*, 57(6): 2714–2725. 2014.

[213] **Catterall, J., Jones, L., and Turner, G.**, Membrane protein glycosylation and CD44 content in the adhesion of human ovarian cancer cells to hyaluronan. *Clinical & Experimental Metastasis*, 17(7): 583–591. 1999.

[214] **Sandmaier, B.M., Storb, R., Bennett, K.L., Appelbaum, F.R., and Santos, E.B.**, Epitope specificity of CD44 for monoclonal antibody–dependent facilitation of marrow engraftment in a canine model. *Blood*, 91(9): 3494–3502. 1998.

[215] **Foley, B.L., Tessier, M.B., and Woods, R.J.**, Carbohydrate force fields. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(4): 652–697. 2012.

[216] **Arsiccio, A., Ganguly, P., La Cortiglia, L., Shea, J.E., and Pisano, R.**, The ADD Force Field for Sugars and Polyols: Predicting the Additivity of Protein-Osmolyte Interaction. *The Journal of Physical Chemistry B*. 2020.

[217] **Guvench, O., Greene, S.N., Kamath, G., Brady, J.W., Venable, R.M., Pastor, R.W., and Mackerell Jr, A.D.**, Additive empirical force field for hexopyranose monosaccharides. *Journal of Computational Chemistry*, 29(15): 2543–2564. 2008.

[218] **Guvench, O., Hatcher, E., Venable, R.M., Pastor, R.W., and MacKerell Jr, A.D.**, CHARMM additive all-atom force field for glycosidic linkages between hexopyranoses. *Journal of Chemical Theory and Computation*, 5(9): 2353–2370. 2009.

[219] **Hatcher, E., Guvench, O., and MacKerell Jr, A.D.**, CHARMM additive all-atom force field for aldopentofuranoses, methyl-aldopentofuranosides, and fructofuranose. *The Journal of Physical Chemistry B*, 113(37): 12466–12476. 2009.

[220] **Sauter, J. and Grafmüller, A.**, Solution properties of hemicellulose polysaccharides with four common carbohydrate force fields. *Journal of Chemical Theory and Computation*, 11(4): 1765–1774. 2015.

[221] **Lay, W.K., Miller, M.S., and Elcock, A.H.**, Optimizing solute–solute interactions in the GLYCAM06 and CHARMM36 carbohydrate force fields using osmotic pressure measurements. *Journal of Chemical Theory and Computation*, 12(4): 1401–1407. 2016.

[222] **Paavilainen, S., Róg, T., and Vattulainen, I.**, Analysis of twisting of cellulose nanofibrils in atomistic molecular dynamics simulations. *The Journal of Physical Chemistry B*, 115(14): 3747–3755. 2011.

[223] **Lowery, J.W., Amich, J.M., Andonian, A., and Rosen, V.**, N-linked glycosylation of the bone morphogenetic protein receptor type 2 (BMPR2) enhances ligand binding. *Cellular and Molecular Life Sciences*, 71(16): 3165–3172. 2014.

[224] **Peiris, D., Spector, A.F., Lomax-Browne, H., Azimi, T., Ramesh, B., Loizidou, M., Welch, H., and Dwek, M.V.**, Cellular glycosylation affects Herceptin binding and sensitivity of breast cancer cells to doxorubicin and growth factors. *Scientific Reports*, 7(43006). 2017.

[225] **Kaszuba, K., Grzybek, M., Orlowski, A., Danne, R., Róg, T., Simons, K., Coskun, Ü., and Vattulainen, I.**, N-Glycosylation as determinant of epidermal growth factor receptor conformation in membranes. *Proceedings of the National Academy of Sciences*, p. 201503262. 2015.

[226] **Hirabayashi, J.**, Lectin-based structural glycomics: glycoproteomics and glycan profiling. *Glycoconjugate Journal*, 21(1-2): 35–40. 2004.

[227] **Wuhrer, M., Catalina, M.I., Deelder, A.M., and Hokke, C.H.**, Glycoproteomics based on tandem mass spectrometry of glycopeptides. *Journal of Chromatography B*, 849(1-2): 115–128. 2007.

[228] **Pan, S., Chen, R., Aebersold, R., and Brentnall, T.A.**, Mass spectrometry based glycoproteomics — from a proteomics perspective. *Molecular & Cellular Proteomics*, 10(1). 2011.

[229] **Waters, M.J. and Brooks, A.J.**, JAK2 activation by growth hormone and other cytokines. *Biochemical Journal*, 466(1): 1–11. 2015.

[230] **Cunningham, B.C., Ultsch, M., De Vos, A.M., Mulkerrin, M.G., Clauser, K.R., and Wells, J.A.**, Dimerization of the extracellular domain of the human growth hormone receptor by a single hormone molecule. *Science*, 254(5033): 821–825. 1991.

[231] **Brooks, A.J., Dai, W., O'Mara, M.L., Abankwa, D., Chhabra, Y., Pelekanos, R.A., Gardon, O., Tunny, K.A.,**

Blucher, K.M., Morton, C.J., Parker, M.W., Sierecki, E., Gambin, Y., Gomez, G.A., Alexandrov, K., Wilson, I.A., Doxastakis, M., Mark, A.E., and Waters, M.J., Mechanism of activation of protein kinase JAK2 by the growth hormone receptor. *Science*, 344(6185). 2014.

[232] **Gent, J., Van Kerkhof, P., Roza, M., Bu, G., and Strous, G.J.**, Ligand-independent growth hormone receptor dimerization occurs in the endoplasmic reticulum and is required for ubiquitin system-dependent endocytosis. *Proceedings of the National Academy of Sciences*, 99(15): 9858–9863. 2002.

[233] **Raivola, J., Haikarainen, T., and Silvennoinen, O.**, Characterization of JAK1 Pseudokinase Domain in Cytokine Signaling. *Cancers*, 12(1): 78. 2020.

[234] **Braunger, J.A., Brückner, B.R., Nehls, S., Pietuch, A., Gerke, V., Mey, I., Janshoff, A., and Steinem, C.**, Phosphatidylinositol 4, 5-Bisphosphate Alters the Number of Attachment Sites between Ezrin and Actin Filaments A COLLOIDAL PROBE STUDY. *Journal of Biological Chemistry*, 289(14): 9833–9843. 2014.

[235] **Levy, G., Carillo, S., Papoular, B., Cassinat, B., Zini, J.M., Leroy, E., Varghese, L.N., Chachoua, I., Defour, J.P., Smith, S.O., and Constantinescu, S.N.**, MPL mutations in essential thrombocythemia uncover a common path of activation with eltrombopag dependent on W491. *Blood*, 135(12): 948–953. 2020.

[236] **Pikman, Y., Lee, B.H., Mercher, T., McDowell, E., Ebert, B.L., Gozo, M., Cuker, A., Wernig, G., Moore, S., Galinsky, I., DeAngelo, D.J., Clark, J.J., Lee, S.J., Golub, T.R., Wadleigh, M., Gilliland, D.G., and Levine, R.L.**, MPLW515L is a novel somatic activating mutation in myelofibrosis with myeloid metaplasia. *PLoS Medicine*, 3(7): e270. 2006.

[237] **Defour, J.P., Itaya, M., Gryshkova, V., Brett, I.C., Pecquet, C., Sato, T., Smith, S.O., and Constantinescu, S.N.**, Tryptophan at the transmembrane–cytosolic junction modulates thrombopoietin receptor dimerization and activation. *Proceedings of the National Academy of Sciences*, 110(7): 2540–2545. 2013.

[238] **Aruffo, A., Stamenkovic, I., Melnick, M., Underhill, C.B., and Seed, B.**, CD44 is the principal cell surface receptor for hyaluronate. *Cell*, 61(7): 1303–1313. 1990.

[239] **Skandalis, S.S., Karalis, T.T., Chatzopoulos, A., and Karamanos, N.K.**, Hyaluronan-CD44 axis orchestrates cancer stem cell functions. *Cellular Signalling*, 63: 109377. 2019.

[240] **Liang, J., Jiang, D., and Noble, P.W.**, Hyaluronan as a therapeutic target in human diseases. *Advanced Drug Delivery Reviews*, 97: 186–203. 2016.

[241] **Naor, D., Nedvetzki, S., Golan, I., Melnik, L., and Faitelson, Y.**, CD44 in cancer. *Critical Reviews in Clinical Laboratory Sciences*, 39(6): 527–579. 2002.

[242] **Bourguignon, L.Y., Gilad, E., and Peyrollier, K.**, Heregulin-mediated ErbB2-ERK signaling activates hyaluronan synthases leading to CD44-dependent ovarian tumor cell growth and migration. *Journal of Biological Chemistry*, 282(27): 19426–19441. 2007.

[243] **Toole, B.P.**, Hyaluronan-CD44 interactions in cancer: paradoxes and possibilities. *Clinical Cancer Research*, 15(24): 7462–7468. 2009.

[244] **Liao, Y.H., Jones, S.A., Forbes, B., Martin, G.P., and Brown, M.B.**, Hyaluronan: pharmaceutical characterization and drug delivery. *Drug Delivery*, 12(6): 327–342. 2005.

[245] **Qhattal, H.S.S. and Liu, X.**, Characterization of CD44-mediated cancer cell uptake and intracellular distribution of hyaluronan-grafted liposomes. *Molecular Pharmaceutics*, 8(4): 1233–1246. 2011.

[246] **Luo, Y., Bernshaw, N.J., Lu, Z.R., Kopecek, J., and Prestwich, G.D.**, Targeted delivery of doxorubicin by HPMA copolymer-hyaluronan bioconjugates. *Pharmaceutical Research*, 19(4): 396–402. 2002.

[247] **Ganesh, S., Iyer, A.K., Morrissey, D.V., and Amiji, M.M.**, Hyaluronic acid based self-assembling nanosystems for CD44 target mediated siRNA delivery to solid tumors. *Biomaterials*, 34(13): 3489–3502. 2013.

[248] **Liu, H.n., Guo, N.n., Guo, W.w., Huang-Fu, M.y., Vakili, M.R., Chen, J.j., Xu, W.h., Wei, Q.c., Han, M., Lavasanifar, A., and Gao, J.q.**, Delivery of mitochondriotropic doxorubicin derivatives using self-assembling hyaluronic acid nanocarriers in doxorubicin-resistant breast cancer. *Acta Pharmacologica Sinica*, 39(10): 1681–1692. 2018.

[249] **Bassi, P., Volpe, A., D'Agostino, D., Palermo, G., Renier, D., Franchini, S., Rosato, A., and Racioppi, M.**, Paclitaxel-hyaluronic acid for intravesical therapy of bacillus Calmette-Guerin refractory carcinoma in situ of the bladder: results of a phase I study. *The Journal of Urology*, 185(2): 445–449. 2011.

[250] **Colnot, D.R., Roos, J.C., De Bree, R., Wilhelm, A.J., Kummer, J.A., Hanft, G., Heider, K.H., Stehle, G., Snow, G.B., and Van Dongen, G.A.**, Safety, biodistribution, pharmacokinetics, and immunogenicity of ^{99m} Tc-labeled humanized monoclonal antibody BIWA 4 (bivatuzumab) in patients with squamous cell carcinoma of the head and neck. *Cancer Immunology, Immunotherapy*, 52(9): 576–582. 2003.

[251] **Börjesson, P.K., Postema, E.J., Roos, J.C., Colnot, D.R., Marres, H.A., Van Schie, M.H., Stehle, G., De Bree, R., Snow, G.B., Oyen, W.J., and van Dongen, G.A.M.S.**, Phase I therapy study with ^{186}Re-labeled humanized monoclonal antibody BIWA 4 (bivatuzumab) in patients with head and neck squamous cell carcinoma. *Clinical Cancer Research*, 9(10): 3961s–3972s. 2003.

[252] **De Bree, R., Roos, J.C., Quak, J.J., Den Hollander, W., Snow, G.B., and Van Dongen, G.**, Radioimmunoscintigraphy and biodistribution of technetium-99m-labeled monoclonal antibody U36 in patients with head and neck cancer. *Clinical Cancer Research*, 1(6): 591–598. 1995.

[253] **Börjesson, P.K., Jauw, Y.W., de Bree, R., Roos, J.C., Castelijns, J.A., Leemans, C.R., van Dongen, G.A., and Boellaard, R.**, Radiation dosimetry of 89Zr-labeled chimeric monoclonal antibody U36 as used for immuno-PET in head and neck cancer patients. *Journal of Nuclear Medicine*, 50(11): 1828–1836. 2009.

[254] **Charrad, R.S., Gadhoum, Z., Qi, J., Glachant, A., Allouche, M., Jasmin, C., Chomienne, C., and Smadja-Joffe, F.**, Effects of anti-CD44 monoclonal antibodies on differentiation and apoptosis of human myeloid leukemia cell lines. *Blood*, 99(1): 290–299. 2002.

[255] **Campisi, M. and Renier, D.**, ONCOFID^{TM}-P a Hyaluronic Acid Paclitaxel Conjugate for the Treatment of Refractory Bladder Cancer and Peritoneal Carcinosis. *Current Bioactive Compounds*, 7(1): 27–32. 2011.

[256] **Tijink, B.M., Buter, J., De Bree, R., Giaccone, G., Lang, M.S., Staab, A., Leemans, C.R., and Van Dongen, G.A.**, A phase I dose escalation study with anti-CD44v6 bivatuzumab mertansine in patients with incurable squamous cell carcinoma of the head and neck or esophagus. *Clinical Cancer Research*, 12(20): 6064–6072. 2006.

[257] **Ayaz, P., Hammarén, H.M., Raivola, J., Sharon, D., Hubbard, S.R., Silvennoinen, O., Shan, Y., and Shaw, D.E.**, Structural models of full-length JAK2 kinase. *BioRxiv*, p. 727727. 2019.

[258] **Kassem, N., Araya-Secchi, R., Bugge, K., Barclay, A., Steinocher, H., Khondker, A., Lenard, A.J., Bürck, J., Ulrich, A.S., Pedersen, M.C., Wang, Y., Rheinstädter, M.C., Pedersen, P.A., Lindorff-Larsen, K., Arleth, L., and Kragelund, B.B.**, Order and disorder–an integrative structure of the full-length human growth hormone receptor. *BioRxiv*. 2020.

[259] **Gadina, M., Le, M.T., Schwartz, D.M., Silvennoinen, O., Nakayamada, S., Yamaoka, K., and O'Shea, J.J.**, Janus kinases to jakinibs: from basic insights to clinical practice. *Rheumatology*, 58(Supplement_1): i4–i16. 2019.

[260] **Virtanen, A.T., Haikarainen, T., Raivola, J., and Silvennoinen, O.**, Selective JAKinibs: prospects in inflammatory and autoimmune diseases. *BioDrugs*, 33(1): 15–32. 2019.

[261] **Roskoski Jr, R.**, Janus kinase (JAK) inhibitors in the treatment of inflammatory and neoplastic diseases. *Pharmacological Research*, 111: 784–803. 2016.

[262] **Verstovsek, S., Mesa, R.A., Gotlib, J., Levy, R.S., Gupta, V., DiPersio, J.F., Catalano, J.V., Deininger, M., Miller, C., Silver, R.T., Talpaz, M., Winton, E.F., Harvey Jr., J.H., Arcasoy, M.O., Hexner, E., Lyons, R.M., Paquette, R., Raza, A., Vaddi, K., Erickson-Viitanen, S., Koumenis, I.L., Sun, W., Sandor, V., and Kantarjian, H.M.**, A double-blind, placebo-controlled trial of ruxolitinib for myelofibrosis. *New England Journal of Medicine*, 366(9): 799–807. 2012.

[263] **Harrison, C., Kiladjian, J.J., Al-Ali, H.K., Gisslinger, H., Waltzman, R., Stalbovskaya, V., McQuitty, M., Hunter,**

D.S., Levy, R., Knoops, L., Cervantes, F., Vannucchi, A.M., Barbui, T., and Barosi, G., JAK inhibition with ruxolitinib versus best available therapy for myelofibrosis. *New England Journal of Medicine*, 366(9): 787–798. 2012.

[264] **Van den Neste, E., André, M., Gastinne, T., Stamatoullas, A., Haioun, C., Belhabri, A., Reman, O., Casasnovas, O., Ghesquieres, H., Verhoef, G., Claessen, M.J., Poirel, H.A., Copin, M.C., Dubois, R., Vandenberghe, P., Stoian, I.A., Cottereau, A.S., Bailly, S., Knoops, L., and Morschhauser, F.**, A phase II study of the oral JAK1/JAK2 inhibitor ruxolitinib in advanced relapsed/refractory Hodgkin lymphoma. *Haematologica*, 103(5): 840–848. 2018.

[265] **Pardanani, A., Gotlib, J.R., Jamieson, C., Cortes, J.E., Talpaz, M., Stone, R.M., Silverman, M.H., Gilliland, D.G., Shorr, J., and Tefferi, A.**, Safety and efficacy of TG101348, a selective JAK2 inhibitor, in myelofibrosis. *Journal of Clinical Oncology*, 29(7): 789. 2011.

[266] **Mesa, R.A., Kiladjian, J.J., Catalano, J.V., Devos, T., Egyed, M., Hellmann, A., McLornan, D., Shimoda, K., Winton, E.F., Deng, W., Dubowy, R.L., Maltzman, J.D., Cervantes, F., and Gotlib, J.**, Simplify-1: A phase III randomized trial of momelotinib versus ruxolitinib in janus kinase inhibitor–naïve patients with myelofibrosis. *Journal of Clinical Oncology*, 35(34): 3844. 2017.

[267] **Hexner, E.O., Serdikoff, C., Jan, M., Swider, C.R., Robinson, C., Yang, S., Angeles, T., Emerson, S.G., Carroll, M., Ruggeri, B., and Dobrzanski, P.**, Lestaurtinib (CEP701) is a JAK2 inhibitor that suppresses JAK2/STAT5 signaling and the proliferation of primary erythroid cells from patients with myeloproliferative disorders. *Blood*, 111(12): 5663–5671. 2008.

[268] **Berdeja, J., Palandri, F., Baer, M., Quick, D., Kiladjian, J., Martinelli, G., Verma, A., Hamid, O., Walgren, R., Pitou, C., Li, P., and Gerds, A.T.**, Phase 2 study of gandotinib (LY2784544) in patients with myeloproliferative neoplasms. *Leukemia Research*, 71: 82–88. 2018.

[269] **Komrokji, R.S., Seymour, J.F., Roberts, A.W., Wadleigh, M., To, L.B., Scherber, R., Turba, E., Dorr, A., Zhu, J.,**

Wang, L., Granston, T., Campbell, M.S., and Mesa, R.A., Results of a phase 2 study of pacritinib (SB1518), a JAK2/JAK2 (V617F) inhibitor, in patients with myelofibrosis. *Blood, The Journal of the American Society of Hematology*, 125(17): 2649–2655. 2015.

[270] **Van Vollenhoven, R.F., Fleischmann, R., Cohen, S., Lee, E.B., García Meijide, J.A., Wagner, S., Forejtova, S., Zwillich, S.H., Gruben, D., Koncz, T., Wallenstein, G.V., Krishnaswami, S., Bradley, J.D., and Wilkinson, B.**, Tofacitinib or adalimumab versus placebo in rheumatoid arthritis. *New England Journal of Medicine*, 367(6): 508–519. 2012.

[271] **Taylor, P.C., Keystone, E.C., van der Heijde, D., Weinblatt, M.E., del Carmen Morales, L., Reyes Gonzaga, J., Yakushin, S., Ishii, T., Emoto, K., Beattie, S., Arora, V., Gaich, C., Rooney, T., Schlichting, D., Macias, W.L., de Bono, S., and Tanaka, Y.**, Baricitinib versus placebo or adalimumab in rheumatoid arthritis. *New England Journal of Medicine*, 376(7): 652–662. 2017.

[272] **Vermeire, S., Schreiber, S., Petryka, R., Kuehbacher, T., Hebuterne, X., Roblin, X., Klopocka, M., Goldis, A., Wisniewska-Jarosinska, M., Baranovsky, A., Sike, R., Stoyanova, K., Tasset, C., Van der Aa, A., and Harrison, P.**, Clinical remission in patients with moderate-to-severe Crohn's disease treated with filgotinib (the FITZROY study): results from a phase 2, double-blind, randomised, placebo-controlled trial. *The Lancet*, 389(10066): 266–275. 2017.

[273] **Papp, K., Menter, M.A., Raman, M., Disch, D., Schlichting, D.E., Gaich, C., Macias, W., Zhang, X., and Janes, J.M.**, A randomized phase 2b trial of baricitinib, an oral Janus kinase (JAK) 1/JAK2 inhibitor, in patients with moderate-to-severe psoriasis. *British Journal of Dermatology*, 174(6): 1266–1276. 2016.

[274] **Bissonnette, R., Papp, K.A., Poulin, Y., Gooderham, M., Raman, M., Mallbris, L., Wang, C., Purohit, V., Mamolo, C., Papacharalambous, J., and Ports, W.C.**, Topical tofacitinib for atopic dermatitis: a phase II a randomized trial. *British Journal of Dermatology*, 175(5): 902–911. 2016.